ResearchNotes

Contents

The URL for reading/downloading issues of *Research Notes* is:
http://www.CambridgeESOL.org/rs_notes

The URL for subscribing to *Research Notes* is:
http://www.CambridgeESOL.org/rs_notes/inform.cfm

Editorial Notes

Welcome to Issue 10 of *Research Notes*, our quarterly publication reporting on matters relating to research, test development and validation within Cambridge ESOL.

2002 has been a particularly busy year for us so far with the introduction of the revised Business English Certificates (BEC) – profiled in *Research Notes 8*, and the new suite of Certificates in English Language Skills (CELS) – profiled in *Research Notes 9*. This year also marks a milestone in the life of the Certificate of Proficiency in English (CPE); CPE is our oldest English language examination and has enjoyed a long and illustrious history since it was first introduced in 1913. In our opening article, Professor Cyril Weir of the University of Surrey, Roehampton, documents critical moments in the history of CPE and shows how changes to this exam reflected changing trends in the world of linguistics and language teaching throughout the 20th century. Rod Boroughs follows this with an article describing in some detail the development of a new task format for the revised CPE Listening paper. A fuller account of development work on all the CPE papers will appear shortly in *Innovation and Continuity: Revising the Cambridge Proficiency Examination*, edited by Cyril Weir and published jointly by UCLES and Cambridge University Press.

Although the CPE Revision Project has largely been completed, other projects to review and revise various aspects of our ESOL examinations continue. Liz Gallivan provides an update in this issue on the proposed changes to the KET/PET Writing papers from 2004. Her article emphasises the important link between what happens in the language teaching/learning context and the way our writing tests are designed, as well as the key role played in the revision process by the consultation with stakeholders, an essential component of all our revision projects. Stuart Shaw maintains the thread on writing test revision in his article on Phase 2 of the project to revise the assessment criteria and scales for IELTS. He reports on three key issues addressed by the IELTS Writing Revision Working Group during Phase 2: the relative merits of analytical and holistic approaches to the assessment of second language writing performance; the definition of appropriate assessment criteria; and the drafting of suitable performance descriptors to form the rating scale.

Our Young Learner English (YLE) tests are also undergoing a review at the present time and Fiona Ball reports on a recent study to investigate the story-telling task in Movers and Flyers. Analysis of a set of story-telling transcripts and responses to an examiner survey provides support for the validity and usefulness of this task, together with suggestions for improvement in task design and examiner training. Also on YLE, Neil Jones reports on a recent study to link the three YLE levels – Starters, Movers and Flyers – within a single framework of reference; this work adds to our understanding of learning gain, facilitates construct validation and helps provide more useful and detailed information to YLE users.

Lynda Taylor discusses the current debate relating to 'world Englishes' and reflects on the importance of this issue for English language test providers. Last but not least, we have included some brief reports of various other activities and projects of interest.

Published November 2002

Innovation and continuity: CPE – past and present

CYRIL WEIR, CONSULTANT, CAMBRIDGE ESOL

Introduction

Language tests from the distant past are important historical documents. They can help inform us about attitudes to language and language teaching when there remains little other evidence of what went on in bygone language classrooms. UCLES Cambridge ESOL's Certificate of Proficiency in English (CPE) has by far the longest track record of any serious EFL examination still in existence today so it is a particularly useful vehicle for researching where we have come from in language teaching and testing.

In this paper[1] we try to piece together the development of the UCLES CPE over the last century. This is no simple task as there had been no conscious effort to document its progress at any stage until recently (see Taylor 1979, Spolsky 1995). By trying to document critical moments in the exam's history we can try to understand the forces that have shaped it. We finish by providing a brief summary of the December 2002 changes to bring the history up to the present day.

CPE 1913–1945

Cambridge's formal entry into testing the English of foreigners was not until 1913, when the Certificate of Proficiency in English (CPE) was instituted by the Local Examinations Syndicate (Roach undated: 5). The examination imitated the traditional essay-based native speaker language syllabus: it included an English literature paper (the same as sat by native speakers for university matriculation), an essay, but also a compulsory phonetics paper with a grammar section, and translation from and into French and German. There was also an oral component with dictation, reading aloud and conversation. In all, the candidates spent 12 hours on an extremely demanding test of their abilities in English.

CPE IN 1913

(i) Written:

a. Translation from English into French or German (2 hours)

b. Translation from French or German into English, and questions on English Grammar (2½ hours)

c. English Essay (2 hours)

d. English Literature (3 hours)

e. English Phonetics (1½ hours)

(ii) Oral:

Dictation (½ hour)

Reading and Conversation (½ hour)

1. This paper is largely based on Chapter 1 of C J Weir and M Milanovic (eds) 2002 *Innovation and Continuity: Revising the Cambridge Proficiency Examination*, UCLES/CUP

The test corresponds closely to the contents of Sweet's *The Practical Study of Languages, A Guide for Teachers and Learners* described by Howatt (1984). It is interesting to note that an oral test (reading aloud and conversation), with associated dictation, was present in an international EFL test at such an early stage. This multi componential approach was to differentiate the UCLES Main Suite examinations from most of its competitors throughout the 20th century.

In 1930 a special literature paper for foreign students was provided for the first time, and, compared to the 1913 exam, the choice of essay topics has become more general. In 1913, the choice was very anglocentric:

1. The effect of political movements upon nineteenth century literature in England.
2. English Pre-Raphaelitism
3. Elizabethan travel and discovery
4. The Indian Mutiny
5. The development of local self-government
6. Matthew Arnold

By 1930, subjects are more general and suitable for the variety of candidates:

1. The topic that is most discussed in your country at the present time.
2. Fascism
3. The best month in the year
4. Good companions
5. Any English writer of the twentieth century.
6. Does satire ever effect its purpose, or do any good?

In the same year plans were laid by Roach to adapt the examination to the needs of a wider public. The regulations for 1932 were published in May 1931; the paper on Phonetics had disappeared as a formal test (and so too the earlier questions on English grammar in the translation paper). By 1938 translation papers were being regularly set in a number of languages and papers in other languages were available on request. Choices (two out of three) were given in July 1938 in the 'From English' translation paper, whereas no choice had been offered in 1923. A 'history' alternative could be offered in lieu of 'literature', as an approach to the study of *English Life and Institutions* – a paper which was introduced under that title in the following year.

CPE IN 1938

(i) Written:

a. English Literature (3 hours)
General Economic and Commercial Knowledge (3 hours)

b. Translation from English (2 hours) – 2 out of 3 passages

c. Translation into English (2 hours) – 2 passages

d. English Composition (2½ hours)

(ii) Oral:

Dictation, Reading and Conversation

Major syllabus changes 1945–75

A new syllabus for CPE was introduced by UCLES in 1945. Language still only has a small part to play in the examination, with literature and translation now of equivalent value. A broad range of pathways through the examination was also possible, e.g. the alternative options to English literature. This was in all likelihood a response to the varying curriculum content of diverse educational systems in existing and former colonies as well as an attempt to maximize candidate numbers.

Further significant changes had taken place by 1953. It became possible to take a 'Use of English' paper as an alternative to 'Translation'. (This new paper has remained, albeit with changed formats, to this day). The new paper started with a reading passage with short answer questions; then came a sentence reformulation task; a task requiring the recombining of sentences into a more coherent paragraph; a task involving knowledge of how punctuation can change meaning; an editing task; a descriptive writing task; and finally a task testing knowledge of affixes. The long history of the Use of English paper in this form partially explains why the current equivalent is apparently so diverse.

CPE IN 1953

(i) Written:

a. English Literature (3 hours)
alternatively a General English Literature Paper was offered for Overseas Centres which were unable to obtain the texts prescribed for the Eng Lit paper.
or Science Texts
or English Life and Institutions
or Survey of Industry and Commerce

b. Use of English (3 hours)
or Translation from and into English

c. English Language (composition and a passage of English with language questions) (3 hours)

(ii) Oral:

Dictation, Reading and Conversation

However, the conflicting demands of a broad choice of options and parallel test reliability are undeniable; they reflect the sometimes diverse pulls of reliability and validity. The cardinal guiding principle for UCLES was validity followed closely by utility. This does not mean they did not seek to achieve reliability, but reliability was not the overriding determinant of what went into the examination.

The approach was to aim for validity and work on reliability, rather than through the single-minded pursuit of objectivity seriously curtail what CPE would be able to measure. A valid test that might not present perfect psychometric qualities was preferred to an objective test which though always reliable might not measure that much of value, e.g. not test speaking or writing.

Developments in the 1960s: the move towards a language-based examination

In the early 1960's we see the beginnings of a critical shift in the Cambridge language testing tradition, namely the gradual separation of language testing from the testing of literary or cultural knowledge. Taylor (1979: 9) notes that:

> *"… in 1953 a Use of English paper was introduced as an alternative to the compulsory translation test for candidates in whose languages there were difficulties in arranging an examination. As a straight alternative to Translation its popularity was to grow steadily until 1966, when a change in the form of the examination made it possible to take, with the compulsory English language paper, both Use of English and Translation, instead of one of these in conjunction with the Literature paper or one of its alternatives".*

As a result of widespread consultation, a new syllabus was proposed which reflected a shift towards a language-based examination. Thus a new form of the examination was introduced in 1966:

> *The development of a semi-objective paper at Proficiency level, systematically testing usage and vocabulary, made it possible from 1966 to take Proficiency, by appropriate choice of alternative papers, as a purely language examination …* (UCLES 1982:1).

CPE IN 1966

(i) Written:

Candidates must offer (a) English Language and *two* other papers chosen from (b), (c), or (d). No candidate may offer more than one of the alternatives in (b).

a. English Language (composition and a passage or passages of English with language questions. The choice of subjects set for composition will include some for candidates who are specially interested in commerce.) (3 hours)

b. Either English Literature (3 hours)
or Science Texts
or British Life and Institutions
or Survey of Industry and Commerce

c. Use of English (3 hours)

d. Translation from and into English (3 hours)

(ii) Oral:

Dictation, Reading and Conversation

As in 1953, candidates still have to take two other papers in addition to the compulsory 'English Language' paper. However, unlike 1953, candidate can choose both 'Use of English' and 'Translation from and into English' as two additional papers, which means they do not have to take anything from (b) 'English

Literature' or its alternatives. In 1953 and 1966, candidates spend a total of 9 hours for the three written tests, plus the time for the oral test.

By 1966 'British Life and Institutions' paper and 'Survey of Industry and Commerce' paper both include a reading passage with questions, which are testing reading comprehension. In 1955 neither of these papers included a reading passage, and simply tested the productive knowledge of the candidates by requiring them to explain, compare, distinguish, and describe things. Again this may be regarded as part of the shift to measuring language rather than subject competence.

In section (b) of the 'Use of English' paper, 3-option multiple choice items are introduced.

The 1975 revision

The 1975 revision saw the examination taking a shape which, in its broad outline, is familiar to the candidate of today. The listening and speaking tests in particular represented major developments on the 1966 revision and echoed the burgeoning interest in communicative language teaching in the 1970s, i.e. an increasing concern with language in use as opposed to language as a system for study. The 1970s saw a change from teaching language as a system to teaching it as a means of communication (see, for example, Widdowson's 1978 *Teaching Language as Communication*). This trend reflected a growing interest in Applied Linguistics at British Universities as the field gained academic respectability through the work of Chomsky, Firth, Halliday and Hymes. It also echoed national developments in foreign language teaching, such as the Nuffield Project which made the teaching of foreign languages in Britain a matter of public concern (Howatt 1984: 274–275).

In the newly added 'Listening Comprehension' paper (30 minutes) candidates listen to four passages and answer a total of 20 multiple choice questions. 'Reading and Conversation' has become the new 'Interview' paper (12 minutes) which still has 'Conversation' and 'Reading aloud', but now includes such tasks as 'prepared talk' and 'providing appropriate responses in given situations'. The writing tasks often have a focused functional slant, e.g. requiring comparison, argument or narrative description as shown in the following example.

> *Either (a) Discuss whether it is possible to solve the problem of pollution without losing too many of the advantages of modern life.*
>
> *Or (b) Compare the position of woman today with their way of life in your grandparents' times, and comment on the difference.*

These contrast sharply with the open ended essay format of earlier times (e.g. 1930):

> Fascism
>
> Good companions
>
> Any English writer of the twentieth century

In addition there is a new 'Reading Comprehension' paper (1¼ hours) using multiple choice questions to test knowledge of vocabulary and usage as well as a second section designed to test ability to read with comprehension.

The increased reliance on multiple choice formats acknowledged the attention international examinations must pay to the demands of reliability. The direct connection between the exam and British culture was broken and a potential source of test bias much reduced.

CPE IN 1975

PAPER 1: Composition (3 hours)
PAPER 2: Reading Comprehension (1¼ hours)
PAPER 3: Use of English (3 hours)
PAPER 4: Listening Comprehension (30 minutes)
PAPER 5: Interview (Approx. 12 minutes)

The five papers have replaced the old division of Oral and Written and indicate some movement to recognising further the need to address the notion that language proficiency is not unitary but partially divisible. It was to take a number of Applied Linguists rather longer to discard their firmly held convictions that language proficiency was unitary and that therefore it mattered little what was tested as long as it was done reliably (see Oller 1979).

The length of the examination has also been reduced to about 8 hours as against 12 hours in 1913, 11 in 1926, 9 in 1953 and 1966. It is difficult to ascertain precisely why this was done as no evidence is available concerning the equal effectiveness of shortened forms or public demand for such.

The 1984 revision

In the 1984 revision an attempt was made to put further clear water between the exam and its literary and cultural heritage. This was not a complete break, though, as in the writing part candidates could still offer an essay based on a set book.

The impression is given of a wide-ranging examination where students with diverse future needs are provided with a broad based picture of their general language proficiency, in terms of both use and usage, i.e. their knowledge of English and their ability to use it in a communicative manner.

CPE IN 1984

PAPER 1: Reading Comprehension (1 hour)
PAPER 2: Composition (2 hours)
PAPER 3: Use of English (2 hours)
PAPER 4: Listening Comprehension (Approx. 30 minutes)
PAPER 5: Interview (Approx. 20 minutes)

The examination has also become more streamlined with total exam time down to less than 6 hours as against 8 in 1975 (and 12 in 1913).

The rationale for the changes and the widespread public consultation are described in detail by UCLES:

> *"From 1984 there were modifications in the structure of the two main examinations, the Certificate of Proficiency in English and the First Certificate in English. The examinations retain the plan*

introduced in 1975, with five compulsory written or oral tests, but were shorter and modified in approach in the light of subsequent experience and the detailed consultation with centres carried out in 1980–81. This consultation covered many aspects of the needs of foreign users of English, teaching courses and examination preparation, and the content and purpose of tests and types of question or activity in the present range of tests. The need for international viability was stressed, and the need to avoid as far as possible without alteration of standard, too close an identification with insufficiently general study goals such as the British-based arts course to which the Proficiency examination is felt to be too closely linked. Strong emphasis was placed by centres, too, on the need to link the form and content of the examination at both levels, even more closely than with the 1975 changes, with communicative approaches in teaching, particularly with regard to the content and weighting of the oral element." (UCLES 1987: 1)

The developments mapped out in this article represent the major changes that took place in the UCLES CPE examinations between 1913 and the last major revision in 1984.

The 2002 revision

The December 2002 CPE revision will continue the work first evidenced in the 1975 and the 1984 revisions. In brief, the revised version of CPE now has:

- **Clearer specifications:** a more comprehensive handbook aimed at familiarizing candidates and teachers with the demands made by the examination. Close linking through the ALTE framework to the Common European framework.
- **A paired speaking test:** after research into the relative effectiveness of the former interview with a single candidate and the paired candidate format it emerged that the latter clearly produced a wider range of functional language use.
- **Interlocutor frame:** the interlocutor frame helps ensure that the language of the interviewer is controlled.
- **Variety of sources in reading and text based tasks:** increase in the number of texts to enhance content coverage.
- **Wider range of tasks in writing:** choice of an increased number of different text types.
- **Wider range of real life contexts in listening:** increase in content coverage.
- **Innovative item types:** in Paper 3 e.g. collocation task.

Continuity and innovation will continue to be the twin pillars upon which CPE is based in this century as it was in the last.

References and further reading

Bereiter, C and Scardamalia, M. (1987): *The Psychology of Written Composition*, Hillsdale, NJ: Lawrence Erlbaum Associates

Howatt, A P R (1984): *A History of English Language Teaching*, Oxford: Oxford University Press

Oller, J W (1979): *Language Tests at School*, Harlow: Longman

Roach, J O (undated): *"My work" with the Local Examinations Syndicate 1925–45, Part I, Origin of the Cambridge Examinations in English*, Personal papers of J O Roach

Spolsky, B (1995): *Measured Words*, Oxford: Oxford University Press

Sweet, H (1899/1964): *The Practical Study of Languages, A Guide for Teachers and Learners*, London: Dent, Republished by Oxford University Press in 1964, edited by R. Mackin

Taylor, C (1979): *An Assessment of the University of Cambridge Certificate of Proficiency in English*, Unpublished MA Dissertation, University of London

University of Cambridge Local Examinations Syndicate (1982): *Cambridge Examinations in English: Changes of Syllabus in 1984*, Cambridge: University of Cambridge Local Examinations Syndicate

University of Cambridge Local Examinations Syndicate (1987): *English as a Foreign Language: General Handbook*, Cambridge: University of Cambridge Local Examinations Syndicate

Urquhart, A C and Weir, C J (1998): *Reading in a Second Language: Process, Product and Practice*, Harlow: Longman

Weir, C J (ed) (2002): *Innovation and Continuity: Revising the Cambridge Proficiency Examination*, Cambridge: Cambridge University Press

Widdowson, H G (1978): *Teaching Language as Communication*, Oxford: Oxford University Press

Redeveloping Part 1 of the CPE Listening paper

ROD BOROUGHS, MAIN SUITE GROUP

Introduction

It is clearly desirable that a test of listening at the level of CPE should require candidates to engage with a broad range of text types, topics, and interaction patterns. One way to increase this range in comparison with the current paper format, without significantly extending the overall timing, was to include a part with short extracts. Reports produced by the Chair of the Listening paper and by a key item writer recommended trialling a range of task types based on short extracts in order to determine which task type would perform most satisfactorily at this level.

Development 1 (1994–1996)

Tasks based on short extracts were commissioned from experienced item writers. The extracts were to be up to 45 seconds in length, with one item per extract; sets of extracts could be either discrete or linked by theme; the task types could be three- or four-option multiple-choice, multiple-matching, note-taking, or open questions. Suggested sources were play extracts, adverts, radio announcements etc; suggested testing focuses were place, situation, function, addressee, topic, content, speaker, feeling, opinion, purpose, relationship. At the editing stage, it became clear

that not all of these task types were feasible – the length of the extracts would not support four-option multiple-choice; the keys of the note-taking task could not be sufficiently constrained; and the open question task would also generate too many acceptable answers unless the untested parts of a full sentence were supplied in the response, e.g.:

1. You hear a radio announcement about a forthcoming programme.

What will the programme reveal?

It will reveal someone's _____ **1**

After editing the commissioned items, it was decided to trial discrete and themed short extracts with an open question task (as above), and with a three-option multiple-choice task, e.g.:

1. You hear a woman talking on a public phone.

What kind of company is she talking to?

A. a car repair company
B. a taxi company _____ **1**
C. a car rental company

Themed short extracts would also be trialled with a two-part multiple-matching task, e.g.:

You will hear five short extracts in which people talk on the subject of shirts.

TASK ONE

For questions **11–15**, choose the phrase **A–G** which best summarises what each speaker is saying.

A. making a strong recommendation
B. making a slight criticism _____ **11**
C. receiving an unhelpful suggestion _____ **12**
D. recalling unhappy memories _____ **13**
E. expressing surprise _____ **14**
F. making a generalisation _____ **15**
G. receiving some useful advice

TASK TWO

For questions **16–20**, choose the main topic each speaker is talking about from the list **A–G**.

A. how shirts should be worn
B. a manufacturing process _____ **16**
C. a shirt-making material _____ **17**
D. fashions in shirt design _____ **18**
E. the choice of patterns _____ **19**
F. ironing shirts _____ **20**
G. a shirt for all occasions

Trialling (Autumn 1994/Spring 1995)

The aims of trialling were:

- to ascertain whether short extracts (discrete or themed) could be made appropriately difficult and discriminating at CPE level;
- to ascertain which task types – multiple-choice, multiple-matching, or open questions – performed most satisfactorily with short extracts at CPE level.

Six trial tests were administered on 420 candidates in ten countries with significant CPE entries (Argentina, Czechoslovakia, Germany, Norway, Poland, Portugal, Spain, Sweden, Switzerland, UK). The following preliminary conclusions were drawn from the results of the trial tests:

- Objective tasks based on discrete and on themed short extracts achieved generally satisfactory difficulty and discrimination values, indicating that short extracts could support CPE-level tasks:

Text type	Task type	Rasch difficulty estimate	Point biserial
Short extracts	3-option multiple-choice	76	0.34
		65	0.39
		61	0.34
		78	0.33
	Multiple-matching	79	0.47
		85	0.38

- Productive tasks performed less satisfactorily than objective tasks with short extracts. Although discrimination values were adequate, the level of difficulty was too high, indicating that attempts to constrain the key had not been successful:

Text type	Task type	Rasch difficulty estimate	Point biserial
Short extracts	Open questions	95	0.28
		89	0.33

Feedback received on the trial tests from the trialling centres was generally positive: the variety of tasks and topics, the level of difficulty, and the speed and clarity of delivery attracted mainly favourable comment. The short extracts with three-option multiple-choice or open questions proved to be the most popular of the new task types, with 95% of respondents judging them to be of an appropriate level for CPE. However, the format of the two-part matching task had caused candidates some confusion and was considered too difficult by 60% of respondents.

Development 2 (Spring 1998)

After reviewing the first round of trialling, the revision team decided to proceed with the three-option multiple-choice task,

which had had both the most satisfactory statistical results and the most positive feedback. It was also decided that the extracts should be discrete, rather than themed, as these would allow a wider range of topics to be covered.

However, while the majority of items from the first trial were of an acceptable level of difficulty, a number of items had proved to be below the level of CPE, and the revision team was concerned that difficulty levels might decrease as the task type became familiar to candidates. It was also felt to be desirable that the extracts should be clearly differentiated from the extracts included in other lower-level Listening papers in the Cambridge ESOL Main Suite; for example, at 45 seconds, the extracts in the first trial were only 15 seconds longer that those appearing in the FCE Listening paper. Consideration was therefore given to increasing the length of the extracts to one minute, to allow for texts of greater propositional density. The downside of this proposal was that, unless the overall timing of the paper was increased – something that was acknowledged to be undesirable – then there would have to be a reduction in the number of items in the test, which would have a consequent negative impact on reliability. The solution suggested at a meeting of the revision team was to write two items per extract.

To determine whether the longer, denser extracts could support two items, the Chair of the Listening paper was commissioned to write ten sample tasks. As in the first commission, the writer was asked to utilise a wide range of testing focuses as it had yet to be determined which would work best at the CPE level. However, it was suggested that at least one of each pair of items should test gist understanding. To ensure a variety of text types, the writer was asked to take extracts from both monologues and texts involving interacting speakers.

Following the successful editing of these extracts, four experienced item writers were commissioned to produce five more extracts each. Feedback from these item writers suggested that the task would be sustainable.

Trialling (Spring/Autumn 1999)

The aims of trialling were:

- to ascertain whether discrete short extracts with two three-option multiple-choice items per extract could be made appropriately difficult and discriminating at CPE level;
- to ascertain whether discrete short extracts from texts involving interacting speakers would perform equally well as extracts from monologues.

Three trial tests were administered on 252 candidates in 13 countries with significant CPE entries (Argentina, Brazil, Bulgaria, Finland, Hungary, Peru, Philippines, Poland, Portugal, Russia, Spain, Switzerland, UK).

The trial test results indicated that short extracts with two three-option multiple-choice items per extract could be made appropriately difficult and discriminating at CPE level. It was also evident that texts involving interacting speakers could sustain two items:

Text type	Task type	Rasch difficulty estimate	Point biserial
Short extracts (monologues and texts involving interacting speakers)	3-option multiple-choice (with 2 items per extract)	74.5	0.48
		77.8	0.28
		74.85	0.30
		80.9	0.24

Feedback on the new task from trialling centres was very positive. It was commented that the task was at the right level for CPE and that it offered an opportunity to increase the range of text types on the paper.

Development 3 (1998–2000)

The new task was unveiled to teachers in a series of teachers' seminars. 200 CPE teachers attended seminars in Spain, Argentina, Brazil, France, Germany, Greece, Portugal and the UK. A substantial majority of respondents (73%) viewed the inclusion of the three-option multiple-choice task on short extracts as a positive development: the task was regarded as appropriate to the level and useful in increasing the range of texts in the paper.

The new task was similarly welcomed at the Chairs' and Principal Examiners' meeting (March 1999), where it was commented that the task would have positive washback since teachers would find it more practical to replicate for classroom activities than tasks based on longer texts. It was also recognised that the length of the extracts would allow for particular emphasis on the testing of global comprehension – a valuable high-level listening skill which is more difficult to test with longer texts, where gist items tend to overlap with other items testing detailed understanding of parts of the text.

Further support for the task was received at the May 1999 Invitational meeting (attended by EFL consultants, academics, representatives of stakeholder organisations such as BC and ARELS, and representatives of EFL publishers), on the grounds that it would allow for a greater coverage of genre, function and topic within the paper.

In all three consultative exercises, opinion was canvassed as to which part of the paper the task should appear in. It was universally agreed that the task should form Part 1 of the paper, so that the shorter texts might act as a lead-in to the long texts in the rest of the paper.

Part 1 of the revised CPE Listening paper: summary

Part 1, comprising eight three-option multiple-choice items based on four discrete extracts of approximately one minute's duration each, enables candidates to engage with a wider range of text types, topics and styles than was possible in the post-1984-revision paper format. It also allows for a wider coverage of testing focus.

The three-option multiple-choice task, being an objective task type, is particularly suitable to the testing of attitude, opinion and inference. However, the full range of testing focuses possible in this part of the paper include general gist, function, speaker/addressee, topic, feeling, attitude, opinion, purpose/intention, genre, course of action, and place/situation. In addition, the length of the extracts allows for particular emphasis on the testing of global understanding. Part 1 also provides a graded first section to the test, whereby candidates have four 'fresh starts', before moving on to the longer texts in Parts 2, 3 and 4. Finally, since short extract tasks are a feature of both the FCE and CAE Listening papers, their inclusion at CPE level helps to give a 'family likeness' across the Main Suite of examinations, as well as maintaining continuity of approach in examination preparation courses.

Update on changes to the KET/PET Writing papers from 2004

LIZ GALLIVAN, MAIN SUITE GROUP

Introduction

Issue 7 of *Research Notes* (February 2002) included an article on the current review of the KET/PET examinations, with particular reference to the consultation process which always constitutes the first stage of any exam review or revision. This article focuses specifically on the changes which are being proposed to the Writing papers for KET and PET from 2004.

Once the initial consultation period was over, proposed changes were only decided after extensive trialling. For example, the changes to the PET Writing paper were trialled, not only to confirm that the content of the tasks was at the right level for PET students, but also to see that candidates were being given enough time to complete the tasks. Students and teachers from many countries helped with the trialling and to refine the tasks, timing and the markschemes. Key changes to the Writing papers for KET and PET are described below.

Key English Test

KET Part 6

This is an example question from the new task in the KET Writing section. The instructions tell the student that all the sentences are about different jobs.

Example:
I help people to learn things. t **e a c h e r**

The objective was to find a way to test candidates' productive knowledge of lexical sets. Our stakeholder survey research shows that students at KET level focus on learning the spelling of vocabulary items, and that in classrooms around the world teachers are encouraging students to keep vocabulary notebooks to store and learn words thematically. This task keeps elements of the old Part 2 task, providing a learner-dictionary style definition of a word in a given lexical set, while adding the productive element of completing a word, having been given the initial letter.

KET Part 8

The information transfer task (previously Part 7) has been amended to make it look more authentic. Candidates will now have to look at up to two, related, input texts in order to extract the information they need to complete a more authentic output text. Examples of this task can be found in the Updated Specifications for KET.

KET Part 9

The focus of the guided writing task has always been on the communicative ability of the candidate and it continues to be so. Feedback from test users was that KET-level candidates frequently produce slightly longer pieces of writing than required in the current exam. In addition, a survey of KET test scripts over the past few years shows that the average candidate produces around 30 words in this part of the test. For this reason, the number of words the candidates need to write has increased from 20–25 to 25–35. This change reflects what is actually happening in English language classes and exam rooms.

Preliminary English Test

PET Part 1

Students will be given the beginning and end of the sentence for their sentence transformation task and will need to fill the space with 1–3 words. This will focus the task solely onto the correct identification of the target structure.

Example:
I prefer playing tennis to playing squash.
I like playing tennis ***more than*** *playing squash.*

PET Part 2

The current form-filling task will be replaced with a guided writing task, with a strong communicative purpose. This will expand the range of text types that PET students produce in the writing component, in line with the feedback we received from schools on what happens in their classrooms.

Example:
An English friend of yours called James gave a party yesterday, which you enjoyed. Write a card to James. In your card, you should,

- *thank him for the party*
- *say what you liked best*
- *suggest when you could both meet again.*

Write 35–45 words on your answer sheet.

The task will be marked with emphasis on how successfully the student communicates the three content elements. The following markscheme, sample scripts and commentaries illustrate how the example question above would be marked.

PET PART 2: MARKSCHEME

5	All content elements covered appropriately. Message clearly communicated to reader.
4	All content elements adequately dealt with. Message communicated successfully, on the whole.
3	All content elements attempted. Message requires some effort by the reader. *or* One content element omitted but others clearly communicated.
2	Two content elements omitted, or unsuccessfully dealt with. Message only partly communicated to reader. *or* Script may be slightly short (20–25 words).
1	Little relevant content and/or message requires excessive effort by the reader, or short (10–19 words).
0	Totally irrelevant or totally incomprehensible or too short (under 10 words).

Sample 1:

Hi, James.

Your party yesterday was very nice. Thanks for inviting me. The best I liked was that interesting game we played. I think we could meet on next Saturday because on Friday I have school.

Bye, Your dear, Ali

5 marks
Commentary: All content elements are covered appropriately. Minor language errors do not impede clear communication of the message.

Sample 2:

Dear James

The party was great! I've never been in so interesting party. Thank you for organising this party! The best in your party were music and attractive games. I want to meet you again.

Your friend, Maria

3 marks
Commentary: One content element is omitted (does not suggest when to meet again). The other content elements are clearly communicated.

PET Part 3

In the updated exam there will be a choice of extended writing tasks. The introduction of choice means that the exam better reflects the range of writing texts that PET-level students are currently producing in the ESOL classroom. Examples of the tasks in PET Writing Part 3 can be found in the Updated Specifications for PET.

The markscheme for this part, reproduced below, was developed with the help of senior external consultants, including Principal Examiners, and in conjunction with the Cambridge ESOL Performance Testing Unit.

PET PART 3: MARKSCHEME
Note: This markscheme is interpreted at PET level and in conjunction with a task-specific markscheme

Band 5 – Very good attempt
- Confident and ambitious use of language
- Wide range of structures and vocabulary within the task set
- Well organised and coherent, through use of simple linking devices
- Errors are minor, due to ambition and non-impeding

Requires no effort by the reader

Band 4 – Good attempt
- Fairly ambitious use of language
- More than adequate range of structures and vocabulary within the task set
- Evidence of organisation and some linking of sentences
- Some errors, generally non-impeding

Requires only a little effort by the reader

Band 3 – Adequate attempt
- Language is unambitious, or if ambitious, flawed
- Adequate range of structures and vocabulary
- Some attempt at organisation; linking of sentences not always maintained
- A number of errors may be present, but are mostly non-impeding

Requires some effort by the reader

Band 2 – Inadequate attempt
- Language is simplistic/limited/repetitive
- Inadequate range of structures and vocabulary
- Some incoherence; erratic punctuation
- Numerous errors, which sometimes impede communication

Requires considerable effort by the reader

Band 1 – Poor attempt
- Severely restricted command of language
- No evidence of range of structures and vocabulary
- Seriously incoherent; absence of punctuation
- Very poor control; difficult to understand

Requires excessive effort by the reader

0 – Achieves nothing
Language impossible to understand, or totally irrelevant to task.

Updated handbooks for both examinations will be available from April 2003. These will contain complete tests for the Reading/Writing and Listening papers, sample Speaking Test materials and all the information needed to prepare students taking the updated KET and PET examinations in March 2004 and beyond.

References and further reading

Van Ek, J A and Trim, J L M (1990): *Waystage 1990*, Cambridge: Cambridge University Press

Van Ek, J A and Trim, J L M (1990): *Threshold 1990*, Cambridge: Cambridge University Press

Revised specifications and sample materials for updated KET and PET can be requested from:

ESOL Information, Cambridge ESOL, 1 Hills Road, Cambridge, CB1 2EU
esol@ucles.org.uk

IELTS Writing: revising assessment criteria and scales (Phase 2)

STUART D SHAW, RESEARCH AND VALIDATION

Introduction

The initial phase of the revision of assessment criteria and rating scale descriptors for the IELTS Writing Modules was reported in Issue 9 of *Research Notes* (August 2002). A principal aim of the revision project is to improve both the reliability and validity of the writing assessment process for IELTS by redeveloping the rating scale. This article, one of a number in the same series, reports on Phase 2 of the project – the Development Phase – which entailed the design and development of the revised rating scale in preparation for trialling and validation.

Phase 1 of the project – Consultation, Initial Planning and Design – involved consultation with a range of stakeholders and was completed in December 2001. The phase highlighted several key issues from the perspective of the assessor, more particularly, individual approaches and attitudes to IELTS Writing assessment, differing domains (Academic and General Training) and differing task genres (Task 1 and Task 2) – all of which provided a valuable focus for the subsequent re-development of existing rating scale criteria.

In general, rating scales attempt to equate examinee performance to specific verbal descriptions (Upshur and Turner 1995). The development and subsequent revision of a rating scale and the descriptors for each scale level are of great importance for the validity of any assessment (Weigle 2002: 109). McNamara (1996) has pointed out that the scale that is used in assessing direct tests of writing should be representative, either implicitly or explicitly, of the theoretical construct underpinning the test. Moreover, the points on a rating scale, according to Bachman, are "typically defined in terms of either the types of language performance or the levels of abilities that are considered distinctive at different scale points" (1990: 36).

Bachman and Palmer claim that theoretical construct definitions, from which rating scales are constructed, may be founded on either the content of a language learning syllabus or a theoretical model of language ability (1996: 212) and further suggest that scale definition comprises two components :

- the particular features of the language sample to be assessed with the scale, and
- the definition of scale levels in relation to the degree of proficiency of these features.

When either constructing or re-constructing a rating scale it is these features that should be uppermost in the developer's mind. The degree of detail given in the scale definition will depend on how the ratings are to be used and how much detail the raters need to be given in order to arrive at reliable, valid ratings.

Current writing assessment approach

Each of the two IELTS writing tasks is assessed independently with the assessment of Task 2 carrying more weight in marking than Task 1. Detailed band descriptors have been developed to describe written performance at each of the nine IELTS bands. These exist in two formats: as three 'profile' or analytical scales for each task: Task 1 – Task Fulfilment (TF), Coherence and Cohesion (CC) and Vocabulary and Sentence Structure (VSS) and Task 2 – Arguments, Ideas and Evidence (AIE), Communicative Quality (CQ) and Vocabulary and Sentence Structure (VSS), and also as a global or holistic scale (i.e. the descriptors for each task are conflated into a single set of band descriptors). Assessors are able to select the global or profile approach according to whether a script has a 'flat' or 'uneven' profile.

Phase 2 – Development

The Development Phase, which began in January 2002, comprised a two-fold approach to re-developing the existing rating scale. Scale (re)construction is generally regarded as an expert and elaborate process, with "the involvement of a great many people" (Lumley 2001: 49). Traditionally, there has been a tendency for rating scales to be a priori measuring instruments (Fulcher 1996: 208), that is, their development has been dependant upon the intuitive judgement of an 'expert'. In conjunction with a team of external 'experts' – academic consultants and senior examiners with a particular interest in Academic Writing – each contributing expert knowledge of advances in applied linguistics, pedagogy, measurement and testing theory, the current assessment criteria and rating descriptors were first, deconstructed and subsequently re-developed. In addition to an a priori approach, however, recent language testing research has suggested a more empirically-oriented approach to the generation of rating scales through the examination of actual scripts and/or operational ratings of writing performance (Shohamy 1990; Milanovic, Saville, Pollitt and Cook 1996). Analysis of samples of actual language performance constituted an integral part of the Development Phase.

Three key revision areas were identified and addressed during the Development Phase:

1) Assessment approach

Two significant issues in the evaluation of direct tests of writing are choice of an appropriate rating scale and establishing criteria based on the purpose of the assessment. Three main types of rating scales are discussed in the composition literature – primary trait scales, holistic scales, and analytic scales – characterised by two

	Holistic Rating Scale	Analytic Rating Scale
Advantages	• appropriate for ranking candidates; • suitable for arriving at a rapid overall rating; • suitable for large-scale assessments – multiple markings (e.g. large script throughput); • useful for discriminating across a narrow range of assessment bands; • multiple scores given to the same script will tend to improve the reliability of assessment of that script.	• more observations – improved reliability; • vast range of writing performances; • norm-referencing discouraged; • greater discrimination across wider range of assessment bands (9 Bands); • provision of a greater control over what informs the impressions of raters; • removal of tendency to assess impressionistically; • provision of more research data/information; • more appropriate for second-language writers as different features of writing develop at different rates.
Disadvantages	• assumes that all relevant aspects of writing ability develop at the same rate and can thus be captured in a single score; • a single score may mask an uneven writing profile and may be misleading for placement; • constitutes a sorting or ranking procedure and is not designed to offer correction, diagnosis, or feedback; • single scores do not permit raters to differentiate features of writing such as depth and extent of vocabulary, aspects of organisation and control of syntax.; • not often readily interpretable as raters do not always use the same criteria to arrive at the same scores.	• time-consuming especially in large-scale testing programmes; • expensive, especially for large-scale testing programmes; • may distort and misrepresent the writing process; • presupposes that raters will be able to effectively discriminate between certain attributes or skills or features, which may not necessarily be the case.

distinctive characteristics: (1) whether the scale is intended to be specific to a single writing task or more generally, to a class of tasks, and (2) whether a single or multiple scores should be awarded to each script. Whilst primary trait scales are specific to a particular writing task, holistic and analytical scales have gained wide acceptance in second language testing and teaching practices particularly when used for grading multiple tasks (Canale 1981; Carroll 1980; Jacobs, Zinkgraf, Wormouth, Hartfiel and Hughey 1981; Perkins 1983). Research has suggested that reliable and valid information gleaned from both holistic and analytic scoring instruments can inform language test developers, language testers and teachers about the proficiency levels of examinees.

Holistic assessment, where raters arrive at a rapid overall rating, involves one or more raters awarding a single score based on the overall impression of a composition as a whole text or discourse according to its general properties.

A major advantage of holistic assessment is that compositions can be scored quickly and therefore less expensively. White, the best known of the holists, argues that holistic scoring "reinforces the vision of reading and writing as intensely individual activities involving the full self" and that any other approach is "reductive" (1985: 33). Several scholars, however, have noted that holistic scoring lacks a demonstrated theoretical foundation (Charney 1984; Gere 1980; Odell and Cooper 1980) and some researchers have questioned the premises on which certain conclusions about the reliability of this approach have been based.

Analytical scoring, where raters are required to target judgements to nominated features or skills of writing, is a method of subjective scoring which involves the separation of the various features of a composition into components for scoring purposes. An analytical scale focuses raters' scoring and thus ensures reasonable agreement among raters to permit a reliable score to be obtained from summed multiple ratings.

The use of analytic scales has two very practical advantages. Firstly, it permits a profile of the areas of language ability that are rated, and secondly, they tend to reflect what raters do when rating samples of language use.

Analytical scoring can lead to greater reliability as each candidate is awarded a number of scores. Furthermore, analytic scoring can allow for more precise diagnostic reporting, particularly in the case where a candidate's skills may be developing at differing rates reflecting a marked profile. Analytical scores can be used for correlational research, exemption, growth measurement, prediction, placement, and programme evaluation. In addition, analytic scores act as useful guides for providing feedback to students on their compositions and to formative evaluation which may be used.

The relative merits of holistic and analytic scales are summarised in the table above.

In making comparisons between analytic and holistic methods for evaluating writing there are reliability-time trade-offs which need to be taken into account. It is not always possible to state whether holistic assessment is more or less valid than analytic assessment as the practical consequences of any such differences

which might occur between the two approaches are not always significant.

The move to analytical scales in the revision of IELTS Speaking was for reasons of consistent examiner focus and multiple observations. In a recent internal study to investigate variability in General Training Writing (O'Sullivan 2001), the performance of the study markers (26 IELTS examiners – all of whom used the profile marking approach) – was correlated with the original markers (who varied in their use of the profile approach according to the current marking guidelines). Inter-correlations varied markedly, with fully profile-marked scripts achieving the highest values. These findings suggest 'that a move to profile scoring will bring with it a higher degree of consistency'.

The benefits of analytical assessment in relation to the IELTS examination – enhanced reliability through increased observations, wide range of writing performances, greater discrimination across wider range of assessment bands (9 Bands), provision of a greater control over what informs the impressions of raters, removal of the tendency to assess impressionistically, active discouragement of norm-referencing and the provision of research data/information – suggest that analytic assessment outweighs any advantages offered by a global approach to assessment.

Examiners are already trained to mark analytically and choose to profile mark if the writing profile on a task appears 'jagged'. Moreover, trainers encourage raters to profile rather than global mark for reasons of thoroughness and consistency and currently some centres routinely profile mark. Consequently, a decision to remove the element of choice by ensuring compulsory profile marking of all tasks would seem a logical step.

2) Assessment criteria

McNamara (1996) has discussed in some detail how language performance assessment has often lacked a clear theoretical base, and this he applies as much to the assessment criteria as to the tasks used.

Issues related to the assessment criteria for IELTS include :

- Are the Assessment Criteria sufficient to amply describe performance?
- Should they be different for Task 1 and Task 2?
- How should the Assessment Criteria be defined?
- Can the existing Assessment Criteria be used for both General Training and Academic Writing Task 1 and Task 2?

Enough similarity in the two writing tasks exists across the Academic and General Training Modules to warrant the use of the same set of assessment criteria for each rather than developing separate criteria; consequently, a revised set of criteria was developed for Task 1 in both Academic and General Training Modules and a separate set developed for Task 2 in both modules.

Five revised criteria for both Modules and both Tasks were produced :

Task Achievement (Task 1)
Task Achievement refers to the quality and adequacy of the task response. The writing is assessed in terms of:

content

- Are the main points covered?
- Are they clearly described?

and *organisation*

- Is the structure of the writing appropriate to the task and to the content?
- Is it logical?

Task Response (Task 2)
Task Response refers to the quality and adequacy of the task response. The writing is assessed in terms of:

content

- Are the main ideas relevant, and are they well elaborated and supported?

position

- Is the writer's point of view clear?
- Is it effectively presented?

and *organisation*

- Is the structure of the writing appropriate to the task and to the writer's purpose?
- Is it logical?

Coherence and Cohesion (Task 1 and Task 2)
Coherence and Cohesion refers to the ability to link ideas and language together to form coherent, connected writing. Coherence refers to the linking of ideas through logical sequencing, while cohesion refers to the varied and apposite use of *cohesive* devices (e.g. logical connectors, pronouns and conjunctions) to assist in making the conceptual and referential relationships between and within sentences clear.

Lexical Resource (Task 1 and Task 2)
Lexical Resource refers to the range of vocabulary that the candidate shows an ability to use, and the precision with which words are used to express meanings and attitudes. Lexical Resource also covers aspects of mechanical accuracy including spelling and appropriacy of vocabulary, key indicators of which include the use of vocabulary of an appropriate register; collocational patterns; accuracy of word choice, and controlled use of word relations such as synonymy and antonymy.

Grammatical Range and Accuracy (Task 1 and Task 2)
Grammatical Range and Accuracy refers to the range, and the accurate and appropriate use of the candidate's grammatical resource, as manifested in the candidate's writing. Key indicators of grammatical range are the length and complexity of written sentences, the appropriate use of subordinate clauses, and the range of sentence structures.

3) Rating scale descriptors

It is widely recognised in the assessment reliability literature that the shared interpretation of rating scale descriptors cannot be assumed and unless rating scale points define clearly differentiated levels or bands, precise interpretation by different audiences will vary and will do so according to "previous experience, unconscious expectations and subjective preferences regarding the relative importance of different communicative criteria" (Brindley 1998: 63). Raters endeavour to make decisions on the basis of common interpretations of the scale contents. Furthermore, this decision-process is intended to be transparent and simple (Pollit and Murray 1996, Zhang 1998); according to Bachman (1990: 36), for the scale to be precise it must be possible for raters to clearly distinguish among all the different levels defined – factors the IELTS Revision Working Group were conscious of throughout the Development Phase.

Primary aims in revising band descriptors were :

1. to deconstruct existing band descriptors for writing into the five revised scales, and
2. to apply current band descriptors to a range of Task 1 and Task 2 Academic and General Training certification scripts in order to identify features of performance across the band levels resulting in separate descriptors for the respective criteria for Academic and General Training.

In principle, the aim was to devise a common scale that would accommodate both General Training and Academic characteristics, any problems associated with shared criteria becoming apparent during the application of the revised marking instrument to the assessment of scripts.

The band descriptors evolved through a succession of iterative drafts and fine tunings, the final form being an amalgamation of expert contributions and regarded widely as a rater-friendly instrument.

Conclusion

Phase 2 of the project was completed to schedule in July 2002. The combined use of quantitative methodologies (application of draft criteria and scales to sample language performance) and qualitative methodologies (insightful and intuitive judgements derived from 'expert' participants) have informed the re-construction of assessment criteria and scales for the IELTS Writing Test. It is hoped that outstanding issues relating to 'off-task' scripts, incomplete responses, handwriting and legibility, memorised scripts, underlength scripts and the use of idiomatic language will be resolved before the Validation Phase of the project. Future issues of *Research Notes* will report in detail both the trialling and validation of the revised scales.

References and further reading

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

Brindley, G (1998): Describing Language Development? Rating Scales and second language acquisition. In Bachman, L F and Cohen, A D (eds) *Interfaces between second language acquisition and language testing research*, Cambridge: Cambridge University Press

Canale, M (1981): Communication: How to evaluate it? *Bulletin of the Canadian Association of Applied Linguistics* 3 (2), 77–94

Carroll, B (1980): *Testing communicative performance*, Oxford: Pergamon

Charney, D A (1984): The validity of using holistic scoring to evaluate writing: A critical overview, *Research in the Testing of English* 18 (1), 65–81

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

Gere, A R (1980): Written composition: Toward a theory of evaluation, *College English* 42, 44–48

Jacobs, H, Zinkgraf, A, Wormouth, D, Hartfiel, V and Hughey, J (1981): *Testing ESL composition*, Rowley, MA: Newbury House

Lumley, T (2001): *The process of the assessment of writing performance : the rater's perspective,* Unpublished PhD Dissertation, Department of Linguistics and Applied Linguistics, The University of Melbourne

McNamara (1996): *Measuring Second Language Performance*. Harlow: Longman

Milanovic, M, Saville, N, Pollitt, A, and Cook, A (1996): Developing rating scales for CASE: theoretical concerns and analyses, in Cumming, A and Berwick, R (eds) *Validation in Language Testing*, Clevedon: Multilingual Matters, pp 15–33

Odell, L and Cooper, C (1980): Procedures for evaluating writing: Assumptions and needed research, *College English*, 42, 35–43

O'Sullivan, B (2001): *Investigating Variability in a Writing Performance Test: A Multi-Faceted Rasch Approach* University of Reading/UCLES Report

Perkins, K (1983): On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability, *TESOL Quarterly* 17 (4), 651–71

Pollitt, A and Murray, N L (1996) What raters really pay attention to. In Milanovic, M and Saville, N (eds) *Performance Testing, Cognition and Assessment. Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem*, pp 74–91, Cambridge: Cambridge University Press

Shohamy, E (1990): Discourse analysis in language testing, *Annual Review of Applied Linguistics*. 11, 115–131

Upshur, J A and Turner, C E (1995): Constructing rating scales for second language tests, *ELT Journal* Vol. 49 (1), Oxford: Oxford University Press

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

White, E M (1985): *Teaching and assessing writing*, San Francisco, CA: Jossey-Bass

Linking YLE levels into a single framework

NEIL JONES, RESEARCH AND VALIDATION GROUP

Introduction

The Young Learner English exams (YLE) are offered at three levels: Starters, Movers and Flyers. Flyers, as the highest of the three, was designed to be of approximately the same level as KET, i.e. Council of Europe level A2 (ALTE Level 1), while Starters is at near-beginner level (CEF level A1 or below).

A previous study to link YLE Flyers to KET used a special research test compiled from sections of Reading tests from these two exams. This was administered to some KET and YLE students. The study provides a link to the Cambridge levels system.

Attention then moved to linking the YLE levels to each other vertically. This was difficult to do using a similar experimental design to the KET-Flyers study, because of the low level of proficiency involved, and the highly controlled range of language used in the tests at each level.

Therefore a different approach was followed, based on identifying candidates who had taken YLE tests at two or more levels. A database of over 60,000 candidate results exists, and a simple search on names and date of birth proved enough to identify a large number of such cases. While a few candidates actually take two tests on the same day or very close in time to each other, most cases extend over several months or years. The performance data contained in the database is limited to the band score for each paper (Reading/Writing, Listening and Speaking). However, this is sufficient to enable an equating of levels to be attempted.

Patterns of entry

Candidates were grouped by the time elapsed between taking the two tests (three-month time periods were used). Table 1 summarizes the data and shows a marked annual cycle, with a large number of candidates taking the next YLE level after 4 quarters (i.e. 12 months).

Method of equating

The mean bandscore on the two exams at each time period was then plotted. Figure 1 illustrates this using the example of Starters and Movers Reading/Writing. The jagged lines show the mean bandscore in the two exams at each time period.

When candidates take both exams at the same time they score about one band higher in the easier exam. When they wait for 6 time periods (18 months) before taking the higher level exam, then they achieve the same bandscore as they did in the exam at the lower level.

The linear trend lines summarise the pattern and allow us to use all the data to contribute to estimating the true difference in level between the exams.

Table 1: Starters-Movers and Movers-Flyers entries over time

Quarters (i.e. 3-month time periods)	No of candidates taking Starters followed by Movers	No of candidates taking Movers followed by Flyers
0	23	22
1	83	82
2	206	133
3	174	156
4	2151	2102
5	138	114
6	52	70
7	18	61
8	188	165

The slope of the lines is interesting: the downward slope of the Starters line shows, as expected, that candidates who do worse in the first exam wait longer before taking the second exam. The upward slope of the Movers line, however, shows that having waited longer, they actually do somewhat better than candidates who take the second exam perhaps too quickly. This pattern was observed for some but not all of the cases studied.

These plots showed us the difference in difficulty in terms of mean bandscores, which we could then use to anchor separate Rasch analyses of response data for each component and level, thus putting all the levels onto a single scale.

Figure 2 illustrates the outcomes for the Reading/Writing papers.

Findings and discussion

High-facility tests like YLE are difficult to equate perfectly. However, the overall picture presented here is interesting and useful. It shows that there is a rough equivalence between moving up one band at the same level and achieving the same band at the next level up. This equivalence could provide a basis for interpreting performance on YLE, and it could inform a small re-alignment of the band thresholds to achieve a simple, coherent and transparent framework.

Figure 1: Candidates taking Starters and Movers Reading/Writing – Mean band over time

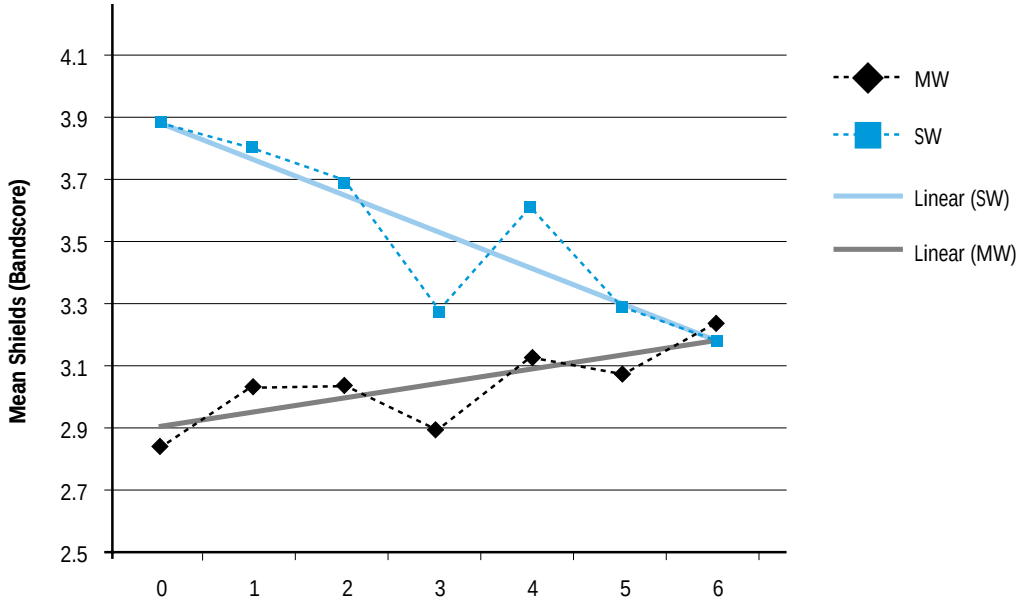
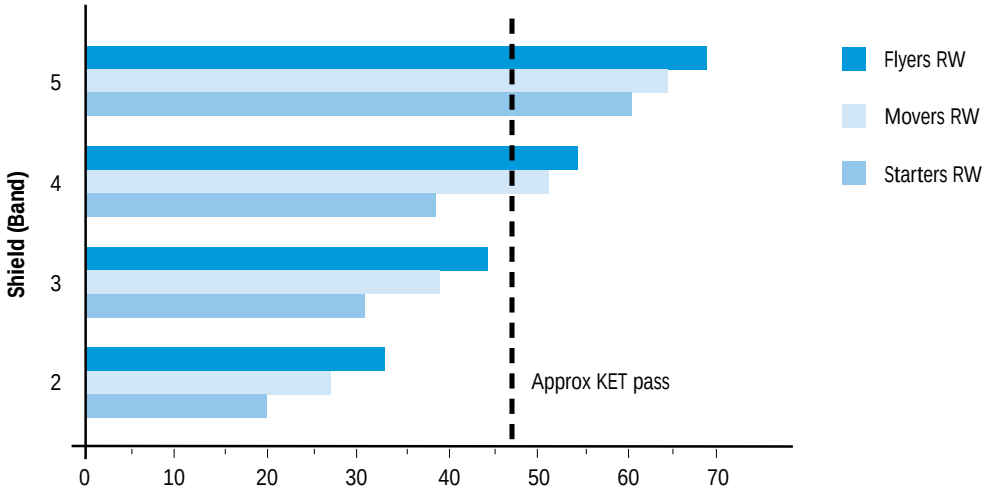


Figure 2: YLE Reading/Writing – 3 levels on the Common Scale



As noted above, there is a tendency to enter candidates for YLE in an annual cycle, which suggests that each level of YLE should be appropriate in difficulty to the typical learning gain being made over this period.

Given the level and the purpose of YLE we should not aim to achieve the kind of psychometric rigour appropriate for exams at higher levels. However, validation of the YLE Framework adds to our understanding of learning gains, facilitates construct validation, and supports the provision of more useful and detailed information to users of YLE.

Language Testing Research Colloquium 2003 – Advance notice

The 25th International Language Testing Research Colloquium (LTRC) will be held at the University of Reading, UK, from 22 to 25th July 2003. Further information is available from the following website: http://www.rdg.ac.uk/AcaDepts/ll/teru/ltrc2003.

Investigating the YLE story-telling task

FIONA BALL, RESEARCH AND VALIDATION GROUP

Introduction

A range of research projects were outlined for the Young Learners Tests in *Research Notes 7* (February 2002). Two projects have recently been completed that relate to the story-telling task in the YLE speaking test. This task consists of candidates narrating a story using a sequence of pictures following a brief explanation of the first picture by the examiner. There has been some concern that story-telling is too difficult a skill for children. This position led to the detailed analysis of ten live YLE speaking test performances and a questionnaire to canvas examiners' views and experiences of the story-telling task. The results of both projects are summarised below.

Qualitative analysis of transcripts

The first project was a qualitative analysis of ten responses to the story-telling task using test performances in Cambridge ESOL's spoken learner corpus currently under development. A qualitative difference was anticipated between the stories produced by Movers and Flyers candidates, possibly attributable to more scope for progression with five pictures than four and more advanced language skills of the higher level candidates. A candidate's familiarity with the task itself was also thought likely to affect their ability to tell a story. Ten speaking tests were chosen to represent a range of candidates in terms of their nationality, gender and age. Each test was transcribed using basic orthography in order to have a written representation of the test.

Two story transcripts are given below to contrast weaker and stronger performances on the same task. Both candidates were given the following introduction by the examiner:

'Now look at these pictures. They show a story. Look at the first one. John's in the garden with his dog. They are playing with a ball. The ball's going into the field. Now you tell the story.'

Candidate A's story

EXAMINER	... now you tell the story
CANDIDATE	erm (1)
EXAMINER	can John find the ball
CANDIDATE	no
EXAMINER	mmhm (1)
CANDIDATE	umm (1)
EXAMINER	what can they see now
CANDIDATE	(1) er like a ball
EXAMINER	mmhm and
CANDIDATE	it was a rabbit
EXAMINER	it was a rabbit thank you very much

Candidate B's story

EXAMINER	... now you tell the story
CANDIDATE	they went out to the field to look for the ball they looked in the grass and everywhere but they can't find the ball
EXAMINER	mmhm
CANDIDATE	after that the saw a black thing in the distance (.) they went after it but it wasn't their ball it was a rabbit
EXAMINER	alright thank you

The whole transcript was read and the story-telling task listened to on cassette before the story-telling task was analysed in three ways:

- the number of words and turns produced by examiner and candidate;
- range and type of language in candidate talk;
- use of linking devices in candidate talk.

There was a wide range in story lengths and the division of turns between examiner and candidate in the ten tests analysed, as shown in the sample stories above. On average, more language was produced at Flyers level than Movers level (53–101 rather than 11–117 words) but the length of examiner contributions was similar at both levels (42–72 words). Some candidates spoke far more than others in the story-telling task, however it must be noted that the better stories were concise whereas candidates who talked a lot tended to hesitate, repeat themselves or to ask questions of the examiner which dramatically increased their word count. The number of turns varied to a similar extent in candidate talk and examiner talk (3–11 and 2–10 turns respectively). The high number of turns in two Flyers stories (10 turns by both examiner and candidate) is probably accounted for by clarification being sought and provided during the story.

The type and range of language exhibited in the stories also varied widely. There were many examples of high level vocabulary and structures and evidence of linguistic awareness in the transcripts, some examples of which are given below:

- Contraction: 'a black thing's jumped up'
- Creativity/generalisation: 'louding' (for shouting loudly)
- Higher level vocabulary: everywhere, go after, distance
- Quantifying: 'two of the cows'
- Reporting: 'I think….'
- Self-correction: 'look with it… look for it'

These examples suggest that YLE candidates have the ability to use certain advanced language skills which this task enables them to display. The combination of unusual vocabulary items in many of the stories and other evidence of language awareness suggest that the story-telling task elicits more complex vocabulary and structures than other parts of the test. The number of interesting features found in ten tests was somewhat surprising although other observations about advanced language should be based on a larger dataset than that analysed here.

The ten candidates also used a wide range and number of linking strategies in their stories. Some candidates used more than ten links whilst others used none. This depended on the interlocutor's support of the candidate using back-up prompts supplied in the interlocutor frame. Some candidates also used links as a 'filler', to give themselves thinking time whilst others used repetition or *um*, *er* to fill pauses in their narrative. Most candidates successfully used a range of linking devices such as *after this/that*, *and now/then*, *as*, *because*, *so*, *but*, *now* or *when*. In some cases the story resembled a question-answer session rather than a story-telling event. The use of back-up questions by examiners will be addressed in another project as the influence of the examiner on candidate performance was revealed from this analysis of speaking tests. Despite this, however, most of the candidates in this sample sounded confident and were eager to tell the story.

All ten candidates managed to tell a story from four or five pictures with reasonable success. Whilst some candidates were stronger on vocabulary, others displayed a particularly clear understanding of the story or linked their discourse consistently. Several candidates gave the researchers a real sense of story-telling which was encouraging and matched some of the positive comments provided by examiners in the follow-up study.

The main conclusions from this project are that the story-telling task is appropriate for the linguistic level of the candidates who take YLE and that there are qualitative and quantitative differences in candidates' production when compared to the other parts of the speaking test. Most candidates produced at least one example of interesting or advanced language use such as a vocabulary item above that level or self-correction. Some candidates used filling strategies to give themselves more preparation time and several candidates asked the interlocutor for clarification whilst telling their story. The story-telling task challenges all candidates and seems to be the component that best distinguishes between weak and strong candidates. It is also the only part of the speaking test that requires candidates to produce a coherent stretch of connected language so it will remain in the YLE Speaking Test at Movers and Flyers levels.

Examiner questionnaire

Following on from the first project a questionnaire on the story-telling task was sent to YLE oral examiners in June 2002. The general response was that this task was a useful means of determining a candidate's level, as higher level candidates give complete stories and lower level candidates tend to describe the pictures or give unlinked sentences. The influence of preparation on performance in the story-telling task was emphasised by many respondents and a quarter of respondents noted that candidates sometimes have difficulty in conceptualising the story. Some examiners noted that candidates get frustrated when they cannot produce a coherent story through lack of vocabulary or understanding. As the vocabulary and structures expected of candidates are covered in the YLE handbook it is hoped that more candidates will be able to tell a coherent story with improved teaching and preparation.

69% of examiners who responded (40 in all) consider the story-telling task to be an effective means of eliciting extended speech from the candidate and the materials were judged to be of high quality with the stories themselves considered to be creative. Almost all respondents (82%) thought that the right number of pictures was provided (four at Movers level and five at Flyers level) although concerns were raised about the cultural specificity of some stories. Cambridge ESOL already addresses this concern by providing examiners with a selection of materials to choose from. The number of sets of YLE materials are currently being increased to improve examiner choice so that the right story can be chosen for each set of candidates.

The majority of respondents view the story-telling task as more difficult than or at the same level of difficulty as other tasks in the speaking test. This view was balanced by several respondents who reported that candidate performance is improving in the story-telling task, perhaps due to better teaching and preparation. Examiners had varying expectations of what candidates would produce in the story-telling task. 34% of examiners expected candidates at both levels to produce a coherent story whilst 66% expected candidates to provide connected sentences. Only 14% expected to hear the correct use of tenses in the candidates' response. 10% of examiners stated that candidates really enjoy the story-telling task, suggesting that it is indeed a suitable and challenging task for all candidates.

Although the majority of respondents were happy with the task itself some areas for improvement were raised which are currently being addressed by Cambridge ESOL. The key issues were how to support weaker candidates and how to reward strong performance which are of equal importance when improving this task.

The result of these two projects is that the story-telling task has been shown to be a valid task in the YLE Speaking Test at Movers

and Flyers levels and will remain part of the speaking test. As part of the YLE Review a number of recommendations for improving the YLE speaking test are currently being considered by Cambridge ESOL, namely:

Materials production

- Move the story-telling task to the penultimate task in Movers (as at Flyers);
- Ensure stories have a simple, coherent storyline;
- Add a title to the picture sheet and number the pictures to help candidates;
- Improve the back-up questions in the interlocutor frame to guide weaker candidates;
- Consider improving the markscheme to reward exceptional performance on this task.

Examiner training

- Improve examiner training, e.g. include linguistic features to watch out for and clearer expectations of candidate performance;
- Provide extra material and advice for examiner trainers e.g. transcripts or tapes of actual tests;
- Explore the possibility of including a YLE section on teachers' websites.

Through these and other measures it is hoped that YLE will continue to remain relevant to young learners of English around the world and to meet their needs more clearly.

Assessing learners' English: but whose/which English(es)?

LYNDA TAYLOR, RESEARCH AND VALIDATION GROUP

Language variation is a well-established and increasingly well-described phenomenon. At the micro-level, variation manifests itself in distinctive features such as phonological, morphological and lexical, syntactic and orthographic aspects; at a more macro-level, it can be seen in discoursal features (to do with rhetorical structure), and pragmatic features (to do with the socio-cultural context of use). The function of language variation is well recognised: it helps to support notions of identity, belonging to a community, being a member of a particular fellowship group. This identity may be regionally based and manifest itself in the form of a particular accent or dialect; or it may be more personally or group based, giving rise to idiolects or sociolects. Linguistic analysis has also identified variation across high/low, formal/informal forms of language, and in recent years we have learned a great deal about the important variations which occur between language in its *spoken* and its *written* forms.

Increasingly sophisticated approaches to linguistic analysis – using discourse and corpus linguistic techniques – have improved our description of language variation and have led to a greater awareness of the issues it raises for the teaching and learning of language, and also for assessment. These include the role and importance of standardisation and norms (where do we get our norms from?), as well as the notion of prescriptivism (does one variety have an inherently higher value than another and should this variety be imposed as widely as possible, cf Received Pronunciation in spoken English?). For the teacher and tester language variation raises practical issues about what to teach – in terms of pedagogy, materials and training, and also what to test – in terms of the standards, norms, models and criteria for judgement we adopt; and the theoretical and practical decisions facing teachers and testers are made even more complicated by the ever increasing pace of language change.

Applied linguists have sought to model the relationships between the varieties of English. One key contributor to the debate has been Braj Kachru who defined *'world Englishes'* as

> *"the functional and formal variations, divergent sociolinguistic contexts, ranges and varieties of English in creativity, and various types of acculturation in parts of the Western and non-Western world"*

In his 1988 work Kachru subdivided varieties of English into 3 categories or circles: *inner*, *outer* and *expanding*; each of these circles relates to the notion of norms in a different way: some are norm- providing or developing, while others tend to be norm-dependent (see Figure 1). A more recent publication – the Longman Grammar of Spoken and Written English (1999) – sets out to evaluate the extent of variation; interestingly, it orders the varieties of English alphabetically to avoid any hint of superiority/inferiority. Contributors to the discussion from within the applied linguistic community have highlighted the recent worldwide increase in L2 speakers of English and have identified a growing number of *'world Englishes'*; others have warned of the danger of linguistic imperialism or have challenged traditional notions of the 'native speaker' model. The current world Englishes debate is one language teachers and language testers cannot ignore.

Testing agencies, in particular, may need to review their traditional positions on setting and evaluating standards in spoken

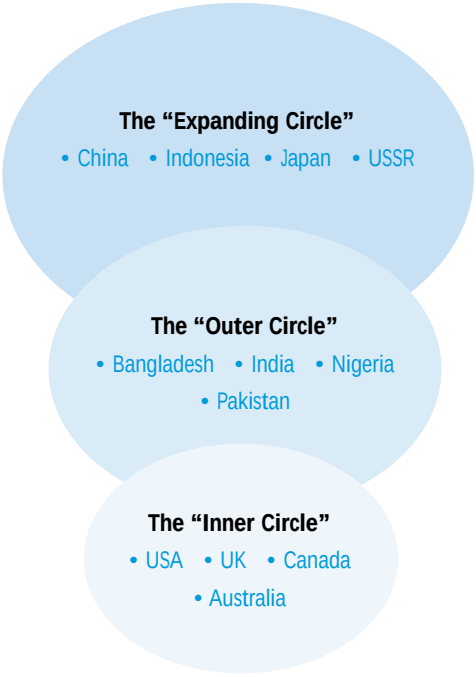


Figure 1: Braj Kachru's circle of World Englishes

and written English. Those of us who are responsible for assessing language ability somehow need to be able to account for language variation within the model of linguistic or communicative competence which underpins our tests; we need to consider how taking account of language variation affects the validity, reliability, practicality and impact of the tests we offer; at the very least we need to keep policy and practice under review and maintain a clear rationale for why we do what we do in relation to the inclusion or non-inclusion of more than one linguistic variety.

It is therefore surprising to discover how little practical discussion there is of this issue in the language testing literature; perhaps it's because examination boards and testing agencies tend to be fairly conservative institutions! Peter Lowenberg is one of the few people to have written on the subject, questioning assumptions about the validity of certain norms. Over recent years he has highlighted numerous examples of test items in commercially produced English language tests which assume a single internationally accepted standard (e.g. standard American English) but which may in fact be biased against English speakers who have grown up with or learnt a different variety of English (e.g. standard British English). For Lowenberg this is an issue of test fairness, a concept which is also much discussed within the language testing community at present and which is especially important where the test is high-stakes, e.g. for gaining access to educational opportunity or career development.

A review of what test agencies say about their stimulus materials, test task design, assessment criteria, standard/norms, and rater training suggests that stated policy and revealed practice vary considerably across test providers when it comes to dealing with world Englishes. Some test providers restrict themselves to 'standard American English' (Michigan) or 'standard North American English' (ETS); others opt for providing alternative test versions (e.g. a British and an American version – LCCI's ELSA), and there are some who appear to make no policy statement at all (Trinity) or who imply they are unbiased towards any variety of standard English (TOEIC).

As a major worldwide provider of English language tests, Cambridge ESOL has been grappling with these issues for some years. We've been providing English language tests since 1913 so historical and developmental factors have clearly shaped our policy and practice over time, but there are also more recent theoretical and practical considerations as the ELT world has changed. One issue is how best to select test input in terms of the content and linguistic features of reading/listening texts and the tasks designed around them. The guiding principles in this case must be to do with the test purpose and the underlying construct, including the need to sample content widely but appropriately without significantly disadvantaging any candidate group. Cambridge ESOL tests reflect the fact that different codes of communication are required by different social contexts; this means that language variation – both social (i.e. formal/informal) and regional – is reflected in the reading and listening texts used in our tests; it is most likely to show up in features of grammar, lexis, spelling, discourse and pronunciation (see Figure 2).

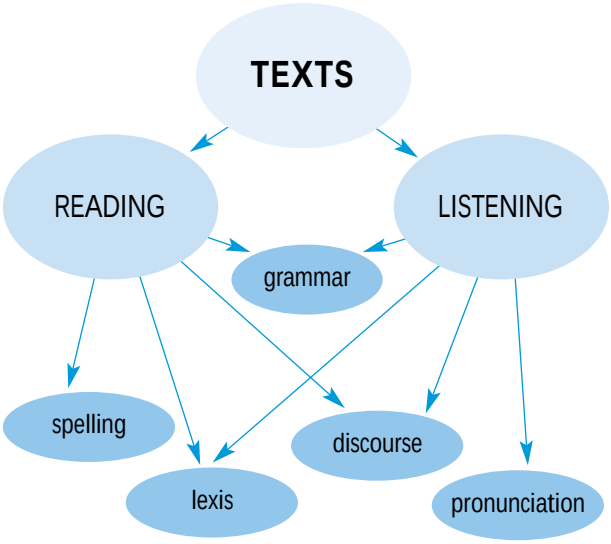


Figure 2: Selecting test input

The notion of the 'dominant host language' can sometimes be a useful guiding principle. For example, in the case of IELTS, which is an international test of English used to assess the level of language needed for study or training in English speaking environments, both texts and test tasks are selected and written by an international team and so reflect features of different varieties of English as the following paragraphs explain:

> *"To reflect the international nature of IELTS, test material is written by trained groups of writers in the UK, Australia and New Zealand."*

"The IELTS Listening Module reflects the fact that different varieties of English are spoken in the contexts around the world in which IELTS candidates are likely to find themselves."

(from *Introduction to IELTS*, p 10)

The nature of the test input is one aspect requiring careful thought; the standards against which we assess a candidate's output are another. How should we treat features of a candidate's written or spoken production (e.g. spelling/pronunciation) which may reflect the variety they have learned or grown up with? Again we have found it helpful to try and articulate a guiding principle:

"Candidates' responses to tasks are acceptable in varieties of English which would enable candidates to function in the widest range of international contexts."

(taken from examination handbooks)

A third key area is the question of who is best qualified to make judgements about standards? Lowenberg suggests that awareness of variation is an essential part of any rater's expertise. In language testing this raises the question of whether NS or NNS examiners are better qualified to evaluate proficiency, and it echoes the wider pedagogical debate about the relative strengths and weaknesses of NS/NNS teachers of English. In relation to Cambridge ESOL, it is certainly not a requirement that a writing or oral examiner for our exams should be a 'native speaker' of English; all examiners (both NS and NNS) are, of course, expected to meet minimum professional requirements in terms of their language competencies.

One could be tempted to ask whether any of this really matters. After all, generally speaking, there is a high degree of uniformity across published English texts around the world. The differences in British and American English standards are restricted to a small set of spelling and lexical variants – and the same may well be true for most other varieties. But it probably does matter – for philosophical as well as pragmatic reasons. One is the changing perceptions about the ownership of English worldwide; another is the increasing rate of language change today. In addition, there is an increasing volume of locally published teaching material and the development of locally generated and defended standards for teaching and learning, especially in certain parts of the world. It also matters because today, more than ever before, we have the appropriate tools to explore the nature of variation across English through corpus-based studies of spoken and written language. For language testers, it matters because we need to be concerned with matters of content and construct validity, and we must pay careful attention to the standard or standards of language we use in our tests. Added to this, we need to take note of the greater focus in today's world on matters of accountability and fairness which impact on professional and public attitudes to tests and test use.

David Graddol has suggested that "the next 20 years will be a critical time for the English language and for those who depend upon it. The patterns of usage and public attitudes to English which develop during this period will have long-term implications for its future in the world." Several experts in the field have speculated on what will become of English, especially in its spoken form; some predict the development of an international language and culture (World Standard Spoken English); others anticipate an increase in regionally-based varieties each with their local cultural conventions and pragmatic norms. But what will this mean for the English that is taught and tested in classrooms and examination rooms around the world? Will it mean the continued teaching and testing of traditionally WASP (white, Anglo-Saxon, Protestant) standards? Or will these be replaced by an international ELF (English as a Lingua Franca)? Or will we see more teaching and testing programs based on one or more standard varieties? What we can predict is that more research will be needed into identifying norms of so-called 'non-native' varieties of English in order to gain a better understanding of how people use English around the world at the start of the 21st century.

At Cambridge we may be ideally placed to make a contribution to this endeavour. The Cambridge Learner Corpus (an electronic database of 15 million words of candidates' writing performance in our tests) could allow us to research questions of interest relating to learner writing and language varieties, e.g.

- Is there evidence in learners' written production of the influence of different varieties of English?
- What is the balance of British and American English usage?
- How is this reflected in learner's use of spelling, lexis, grammar, etc?

The development of our spoken language corpus might in time enable us to investigate the language and behaviour of NS and NNS oral examiners, to explore questions such as:

- How does the language competence of the examiner impact test delivery?
- How does the fact that the examiner and the candidate share an L1 impact test delivery?
- How does the observable behaviour of the NNS examiner resemble that of a comparable NS examiner in the assessment of similar candidates?

As we undertake studies of this nature, we hope to report the findings in future issues of *Research Notes*.

References and further reading

Graddol, D (1997): *The Future of English?* London: British Council

Kachru B B (1992): *The Other Tongue: English Across Cultures* (2nd edition), Urbana: University of Illinois Press

Lowenberg, P H (1993): Issues of validity in tests of English as a world language: whose standards? *World Englishes*, 12/1, 95–106

Lowenberg, P H (2000): Non-native varieties and issues of fairness in testing English as a world language. In A J Kunnan (ed) *Fairness and validation in language assessment, Selected papers from the 19th LTRC, Orlando, Florida*. Studies in Language Testing, Volume 9, UCLES/CUP

Exploring issues in the assessment of pen-and-paper/ computer-based IELTS Writing

For some years now, Cambridge ESOL has been working to develop a computer-based version of IELTS. This has involved a series of research and validation studies. This year we have been able to collaborate with Russell Whitehead of Birkbeck College, London, to explore some of the issues in the assessment of pen-and-paper and computer-based writing. Russell's study was carried out as part of Round 7 of the BC/IDP joint-funded research programme (see *Research Notes 8*) and we were able to supply him with the dataset for analysis.

The study set out to investigate whether candidates taking IELTS Academic Writing tests in computer-based mode would receive the same marks as in pen-and-paper mode, and whether examiners would approach the assessment of computer-based scripts in the same way as for pen-and-paper scripts.

A sample of 50 candidates' scripts and brief questionnaires were collected from six centres which had been involved in the 2001 trialling phase of computer-based IELTS. Candidates in the 2001 trial took a CB version of IELTS followed soon afterwards by their live pen-and-paper IELTS; this meant that for each candidate a pen-and-paper and a computer-generated writing reponse was available for analysis. For Russell's study, six trained and certificated IELTS examiners were recruited to mark approximately 60 scripts each; these consisted of pen-and-paper scripts, computer-based scripts and some pen-and-paper scripts typed up to resemble computer-based scripts. The examiners for the study also completed a questionnaire about the scripts, the assessment process and their experiences of, and attitudes to, assessing handwritten and word processed performance.

Theoretical studies of writing and testing suggest that there are important differences between writing by hand and word processing. However, when the examiners' marks were subjected to Rasch analysis based on the overlapping script allocations, no significant differences were found in marks awarded to candidates in the two modes. Nonetheless, while the current bands are thus 'safe', certain minor aspects of the statistics and a number of comments from candidates and examiners, in conjunction with theory and some previous studies, suggest that a subsequent phase of research should be conducted using protocol studies with candidates and examiners to explore further the processes involved in writing and writing assessment.

Additional trialling of CB IELTS is scheduled to take place over the coming year in several key test centres worldwide and further studies to investigate the issues associated with writing assessment will be built into the research programme.

Lexicom@ITRI: a Lexicography Course

Cambridge ESOL develops and uses corpora for a variety of purposes within the field of language testing. In order to gain insights into corpus use in related fields a member of the Research and Validation Group recently attended a unique training course on lexicography and lexical computing at the Information Technology Research Institute (ITRI) at the University of Brighton. The *Lexicom@ITRI* course involved forty publishers, translators, lexicographers, terminologists and postgraduate students with an interest in dictionary writing and related computing applications. The course was led by two experienced lexicographers (Sue Atkins and Michael Rundell) and a lexical computing expert (Adam Kilgarriff). The latter's British National Corpus wordlists have already been used by Cambridge ESOL as part of the development of wordlists for our item writers to use when developing test materials (see *Research Notes 8*).

Many aspects of dictionary writing were covered during the course including building and exploiting corpus resources, managing dictionary projects and a debate about the future of the printed dictionary. Various lexical computing programs, were demonstrated and delegates were given the opportunity to describe their own corpora, dictionary projects and corpus querying software. Part of the week was spent creating dictionary entries which were presented and discussed at the end of the course. The most relevant contribution to Cambridge ESOL's corpus-related activities was the demonstration of Word Sketch and WASPS software designed by Adam Kilgarriff and colleagues at ITRI. Word Sketch provides snapshots of the grammatical behaviour of words and is an exciting development for lexicography as it reduces the time spent in analysing raw corpus data. Such software could have implications for Cambridge ESOL's own use of corpora in the future as it is hoped that similar software could be used to analyse learner data.

References

Course website:
http://www.itri.bton.ac.uk/lexicom/

Word sketch site:
http://www.itri.bton.ac.uk/~Adam.Kilgarriff/wordsketches.html

Review of recent validation studies

The Research and Validation Group plays a key role in monitoring test performance. As well as analysing candidate performance, we carry out various studies to investigate the performance of oral examiners and test materials – both of which are traditionally regarded as potential sources of variability, especially in the assessment of speaking and writing.

Monitoring oral examiner performance in FCE

Candidates' speaking skills are usually assessed in the Cambridge ESOL examinations using a face-to-face, paired format, i.e. two candidates and two examiners. The advantages of the paired format have been discussed in previous issues of *Research Notes* (see Issue 6). Of the two examiners, one (the Assessor) observes the spoken interaction and rates the candidates' ability using analytical scales – Interactive Communication, Discourse Management, Pronunciation, and Grammar and Vocabulary; the other examiner (the Interlocutor) manages the spoken interaction and rates the candidates' ability using a holistic (or Global Achievement) scale. Cambridge ESOL invests heavily in the training, co-ordination and monitoring of oral examiners worldwide to ensure that quality of assessment is assured and maintained from one test session to another. One of the QA monitoring issues in the Speaking test concerns the level of agreement between Assessor and Interlocutor in their assessment of candidates. The differing roles of the two examiners and their perspectives on the speaking test event are likely to result in a certain level of divergence; nevertheless, there must be an acceptable level of agreement in their ratings since they are using similar assessment criteria.

A recent longitudinal study compared the ratings of nearly 4000 oral examiners who were involved in marking the FCE Paper 5 (Speaking) component during sessions in June 2000 and in June

Table 1: Correlations between Assessor and Interlocutor markings: Syllabus 100

FCE Syllabus 100 June 2000 Paper 5	IC	DM	PR	GV	Average Assessor Mark	GA
Speaking – Interactive Communication Mark (Assessor)	1					
Speaking – Discourse Management Mark (Assessor)	0.82	1				
Speaking – Pronunciation Mark (Assessor)	0.68	0.72	1			
Speaking – Grammar & Vocabulary Mark (Assessor)	0.76	0.86	0.73	1		
Speaking – Average Assessor Mark	0.90	0.94	0.86	0.93	1	
Speaking – Global Achievement Mark (Interlocutor)	0.75	0.78	0.67	0.76	**0.82**	1
Number of Candidates	30475	30475	30475	30475	30475	30475
Number of Examiners (Interlocutors & Assessors)	1864	1864	1864	1864	1864	1864

FCE Syllabus 100 June 2002 Paper 5	IC	DM	PR	GV	Average Assessor Mark	GA
Speaking – Interactive Communication Mark (Assessor)	1					
Speaking – Discourse Management Mark (Assessor)	0.83	1				
Speaking – Pronunciation Mark (Assessor)	0.68	0.72	1			
Speaking – Grammar & Vocabulary Mark (Assessor)	0.76	0.86	0.73	1		
Speaking – Average Assessor Mark	0.90	0.94	0.86	0.92	1	
Speaking – Global Achievement Mark (Interlocutor)	0.75	0.78	0.67	0.76	**0.81**	1
Number of Candidates	36632	36632	36632	36632	36632	36632
Number of Examiners (Interlocutors & Assessors)	1957	1957	1957	1957	1957	1957

2002. The initial findings (see Table 1 opposite) show an acceptably high level of agreement between the marks given by the Assessor and those awarded by the Interlocutor.

The relative stability of correlations between the overall marks given by the examiners over two years is evidence of the effectiveness of training, co-ordination and monitoring of Oral Examiners.

Monitoring IELTS test performance in 2001

Each year, new versions of each of the six IELTS modules are released for use by centres testing IELTS candidates. In addition to the validation work necessary to produce new versions of IELTS, the Research and Validation Group are responsible for estimating and reporting test reliability.

For the Listening and Reading tests this is done using Cronbach's alpha, a reliability estimate which measures the internal consistency of an item-based test. Listening and Reading material released during 2001 had sufficient candidate responses to estimate and report meaningful reliability values as shown below:

IELTS Modules	Alpha
Listening Version A	0.88
Listening Version B	0.85
Listening Version C	0.87
Listening Version D	0.88
Listening Version E	0.89
Academic Reading Version A	0.87
Academic Reading Version B	0.85
Academic Reading Version C	0.84
Academic Reading Version D	0.83
Academic Reading Version E	0.85
Academic Reading Version F	0.87
General Training Reading Version A	0.85
General Training Reading Version B	0.80
General Training Reading Version C	0.86
General Training Reading Version D	0.83
General Training Reading Version E	0.83

The figures reported for Listening and Reading modules indicate expected levels of reliability for tests containing 40 items. Values for the Listening are slightly higher than those for the Reading components; both Academic and General Training candidates take the same Listening module and so the test population reflects a broader range of ability.

Reliability of the Writing and Speaking modules cannot be reported in the same manner because they are not item-based tests; Writing and Speaking modules are assessed at the test centre by qualified and experienced examiners according to detailed descriptive criteria. Reliability of assessment is assured through careful design of the test tasks and assessment scales as well as through the face-to-face training and certification of examiners; all examiners must undergo a re-certification process after two years.

Continuous monitoring of the system-wide reliability of IELTS Writing and Speaking assessment is achieved through a sample monitoring process. Selected centres worldwide are required to provide a representative sample of examiners' marked tapes and scripts such that all examiners working at a centre over a given period are represented. The tapes and scripts are then second-marked by a team of IELTS Senior Examiners. Senior Examiners monitor for quality of both test conduct and rating, and feedback is returned to each centre. Analysis of the paired, examiner-Senior Examiner ratings from the sample monitoring data for 2001 produces correlations of 0.85 for the Writing module and 0.92 for the Speaking module.

The performance of test materials in the Writing and Speaking modules is routinely analysed to check on the comparability of different test versions. Mean Band Scores for the Academic Writing versions released in 2001 ranged from 5.33 to 5.86. Likewise Mean Band Scores for the General Training Writing versions released in 2001 ranged from 5.38 to 5.85. The Mean Band Scores for the Speaking tasks released in 2001 ranged from 5.80 to 5.92. The analysis for both Writing and Speaking shows a very consistent pattern across different test versions and over time.

You can find more information on recent IELTS test performance in the newly published *IELTS Annual Review 2001–2002*, available from any of the IELTS partners.

Monitoring speaking test materials for Young Learners Tests

At each level of Young Learners – Starters, Movers and Flyers – there are six sets of speaking materials (speaking test packs) available for use by examiners during a 12-month period. As with all Cambridge ESOL material, the YL speaking test packs are produced to conform to strict guidelines to ensure that each pack is at the same level of difficulty. Analyses were run on the 2001 speaking packs to provide quantitative evidence that this was happening.

An analysis was carried out on data samples of 837 Starters candidates, 793 Movers candidates and 551 Flyers candidates. The data were split by speaking test pack and mean scores (on a scale of 1 to 6) were calculated for each speaking pack at each level.

Results showed that the variation between the lowest and highest mean score for Starters speaking test packs was 0.26, i.e. approximately a quarter of one mark. The figure for Movers was 0.27 and for Flyers 0.31. In essence this means that for all three YL levels, the speaking test packs appear to perform in very similar ways and there is no evidence to suggest that candidates are likely to score better on one pack than another.

Other news

Common European Framework

One of the most frequent questions asked about any exam is how the levels compare with other qualifications. This is a complex and difficult question, and it is rarely possible to give a straightforward answer, but the most useful comparison is with the Common European Framework, published by the Council of Europe. More information on the Framework and how it relates to our exams is available from ESOL Information: esol@ucles.org.uk

The ALTE (Association of Language Testers in Europe) examinations are the only certificated examinations referred to in the Framework as being specifically anchored to the framework by a long term research programme.

COTE Revision

COTE, the teacher training award which provides professional development for in-service English language teachers, has been revised. The new award will be known as ICELT (In-Service Certificate in English Language Teaching) and is already being offered at some test centres.

A new feature of the revised award is the opportunity for candidates to enter for a separately certificated Language for Teachers module. For more information contact Monica Poulter: poulter.m@ucles.org.uk

Revised CPE

June 2002 was the last administration of the CPE in its current format, and the revised exam will be taken for the first time in December. Support materials available from Cambridge for the revised CPE include the handbook and Upper Main Suite Speaking Test video pack (available for purchase), as well as a leaflet giving details of books and other resources published by CUP, Macmillan, Oxford and Pearson which relate directly to the revised CPE.

Recognition in the USA

We have now published more than 25 information sheets listing institutions that recognise Cambridge ESOL exams and IELTS. One of the most popular of these covers the United States, where a rapidly growing number of universities recognise CAE, CPE and/or IELTS for admissions purposes. International students interested in studying in the USA are encouraged to consult the list, which can be downloaded from our website, and also to contact these institutions directly. If you have questions about recognition in the USA, please contact exams@ceii.org

Shelf life of certificates

We are sometimes asked how long the Cambridge ESOL certificates last, or whether a candidate who took an exam some years ago needs to retake the exam.

The simple answer is that the certificates do not expire. They show that on a particular date the holder demonstrated that they had attained the specified level of language skills. For most candidates, the certificate is the result of a specific preparation course and serves as a mark of achievement in completing the course successfully.

It is clear, however, that language skills can diminish over time – a phenomenon often referred to as 'language attrition'. In deciding whether to rely on a certificate obtained some years ago, educational institutions and employers need to take into account a number of factors, most importantly whether the holder has kept up his or her use of the language and whether the level of the certificate is significantly higher than that required for the job or course in question.

There are therefore no hard-and-fast guidelines for the period since obtaining a Cambridge ESOL certificate after which additional evidence of current language ability may be required by employers or institutions.

The Test Report Form provided by IELTS is not a certificate since it is not focussed on a particular level of language ability; for this reason, the normal shelf life for an IELTS Test Report Form is two years (see under Results in the *IELTS Handbook*).