

# ResearchNotes

## Editorial Notes

Welcome to Issue 8 of *Research Notes*, our quarterly publication reporting on research, test development and validation activity within UCLES EFL.

To coincide with the introduction in the first half of this year of the revised Business English Certificates (BEC) and the revised Standard Test for the Business Language Testing Service (BULATS), this issue contains a special focus on our tests assessing business English. In a useful background article, Barry O'Sullivan discusses some theoretical perspectives on the testing of language for specific purposes, including the testing of language for business. After reviewing the relevant literature, he suggests that tests can be developed along a continuum of specificity (from unspecified to highly specified) and explains how a test placed somewhere in between the extreme ends of the continuum will have the potential to be either more or less generalisable. Neither BEC nor BULATS have been designed as tests which focus narrowly on a particular 'purpose' area where the language and the 'purpose' are both considered as part of the construct; instead, BEC and BULATS are specific purpose language tests in as much as they are located within a business employment context and so reflect linguistic features (lexical, syntactic, etc) of that context. Follow-up articles in this issue describe both tests in more detail: David Booth outlines recent changes to the BEC speaking tests, highlighting key issues which were considered during the revision process and explaining the main outcomes. Ed Hackett reports on the process of revising the BULATS Standard Test and explains the rationale for changes to the Listening, Reading, and Grammar/Vocabulary sections.

Vocabulary lists have always played an important role in the development of many of the UCLES EFL tests, especially those at lower proficiency levels; Fiona Ball describes the ongoing development and validation of the vocabulary lists used in developing the BEC examinations, particularly BEC Preliminary, and she explains the increasing role being played by the Cambridge Learner Corpus in this work.

In an introductory article in *Research Notes 6* Stuart Shaw highlighted some key issues in assessing second language writing and he identified rater training and standardisation as essential to the valid and reliable measurement of writing ability. In this issue he describes a recent experimental study, carried out within the context of the CPE Revision Project, to investigate the effect of the training and standardisation process on rater judgement and inter-rater reliability. This study is part of a much larger and ongoing research programme to deepen our understanding of rater behaviour and to refine our existing approaches to rater training and standardisation.

As well as a short report on recent conferences attended by representatives of the Research and Validation Group, Issue 8 also includes a summary of various validation studies carried out recently which illustrate the broad range of activities undertaken by staff within the Group.

Finally, we include a short report on the joint-funded research programme for IELTS listing the many studies which have been undertaken since the programme was first established in 1995; this issue of *Research Notes* also includes the 2002 call for proposals (Round 8) which we hope will reach as wide an audience as possible.

## Contents

Editorial Notes	1
Some theoretical perspectives on testing language for business	2
Revising the BEC speaking tests	4
Revising the BULATS Standard Test	7
Developing wordlists for BEC	10
The effect of training and standardisation on rater judgement and inter-rater reliability	13
Review of recent validation studies	18
Other news	19
Conference reports	20
Studies in Language Testing – Volume 15	20
IELTS joint-funded research 2002 (Round 8): call for proposals	21
IELTS joint-funded research programme – 1995-2001	22

Please remember to complete our reader questionnaire from *Research Notes 7* to let us know your views on the content and format of the newsletter.

The questionnaire can be accessed and completed on line at:

[http://www.cambridge-efl.org/rs\\_notes/feedback.cfm](http://www.cambridge-efl.org/rs_notes/feedback.cfm)

# Some theoretical perspectives on testing language for business

BARRY O'SULLIVAN, CONSULTANT, UCLES EFL

In the only serious attempt to date to build a theoretical rationale for the testing of language for specific purposes (LSP), Douglas (2000) argues that a theoretical framework can be built around two principal theoretical foundations. The first of these is based on the assumption that language performance varies with the context of that performance. This assumption is rationalised by a well-established literature in the area of sociolinguistics (for example Labov's famous 1963 study) in addition to research in the areas of second language acquisition (Tarone, 1985, 1988, 1998; Tarone & Liu, 1995) and language testing (Berry, 1996, 1997; Brown, 1998; O'Sullivan, 1995, 2000, 2002; Porter, 1991a, 1991b, 1994; Porter & Shen, 1991). In the case of the second foundation, Douglas sees LSP tests as being 'precise' in that they will have lexical, semantic, syntactic and phonological characteristics that distinguish them from the language of more 'general purpose' contexts. This aspect of Douglas' position is also supported by an ever-increasing literature, most notably in the area of corpus-based studies of language in specific contexts (the Wolverhampton Business English Corpus for example).

Douglas places these two foundations within a single over-riding concept, that of authenticity. He defines a test of specific purposes as

*one in which test content and methods are derived from an analysis of a specific purpose target language use situation, so that test tasks and content are authentically representative of tasks in the target situation, allowing for an interaction between the test taker's language ability and specific purpose content knowledge, on one hand, and the test tasks on the other. Such a test allows us to make inferences about a test taker's capacity to use language in the specific purpose domain. (2000:19)*

Douglas (2001) himself acknowledges that there are a number of issues left unanswered by his definition, an argument also made by Elder (2001). This criticism focuses on what Elder (2001) sees as the three principal problematic areas identified in the work of Douglas: the distinguishability of distinct 'specific purpose' contexts; authenticity; and the impact (and interaction) of non-language factors.

## Specificity

While the arguments presented by Elder and Douglas appear to be quite strong, it should be noted that the tests they use to exemplify this argument are not necessarily representative of the genre, in that the degree of specificity is clearly high, and the degree of generalizability demanded is conversely quite low. One aspect of Elder's (2001: 154) criticism of LSP tests focuses on the difficulty of being able to specify the functions to be addressed in such a test.

She cites the infinite possibility of the range of what Swales (1990: 52) called 'allowable contributions'. The degree to which it is possible, or desirable, to attain a high level of specificity in LSP tests is clearly an issue in urgent need of empirical examination. A small number of tests have been developed for domains that are limited to ritualised or routinised language, such as the test for air-traffic controllers reported by Teasdale, 1994. While it may be possible to identify a detailed (and possibly close to complete) range of language tasks associated with a domain, it is never going to be possible to identify the full range of 'allowable contributions'. It may therefore be useful to examine how existing LSP tests deal with this problem.

## Authenticity

While Douglas (2000) bases his definition of LSP tests on the twin aspects of authenticity (situational and interactional), it is still seen, both by Elder (2001) and Douglas (2001), as being one of the principal areas of concern. The notion of situational authenticity is relatively easy to conceptualise. Tests such as that described by Teasdale (1994) where candidates were tested in a situation that closely replicated the specific purpose domain (they wore headphones to respond to 'pilots' under their care) are as close as we can get to a completely situationally authentic test. The mere fact that the event is being used as a test lessens the authenticity.

In the case of interactive authenticity, the situation is quite different. Here, there is a lesser degree of certainty, in that it is not at all clear that all administrations of even one particular test will result in equivalence of input (or output) by the participants. This view has been at least in part addressed by the move on the part of examination boards (particularly by UCLES) towards a more careful monitoring of performance test administration through the introduction of 'interlocutor frames' to control test input, and towards a monitoring of test (and tester) behaviour through direct observation of either live test events or recordings (audio or video) of those same events. Elder's point, however, is one that cannot simply be ignored. While it is relatively straightforward to establish the situational authenticity of a test task, it is only by *a priori* empirical exploration of test performance that evidence of the interactional authenticity of any test can be established.

## Impact of non-language factors

In an interesting argument, in which he advocates a move towards an integrated language/specific area ability approach, Douglas (2000) suggests using what he refers to as 'indigenous' scales (a term first suggested by Jacoby, 1998) in LSP tests. The argument is that the criteria actually employed in the evaluation of specific

purpose performances are specific to the context of that performance – a position which is seen as a retort to the *inseparability* of language and performance of specific purpose task (Elder, 2001; Douglas, 2001). While the case made by Douglas is strong, there are a number of points on which he can be taken to account.

The central problem here is one of construct definition, and therefore of the inferences that are to be drawn from a particular test. The Occupational English Test (OET), for instance, is criticised by Douglas and by its principal creator, McNamara (in Jacoby & McNamara, 1999) for using a ‘general purpose’ rating scale, rather than one devised from an analysis of the target language use (TLU) situation.

However, we should remember that the test, for whatever reason (the one suggested was bureaucratic expedience), was meant to offer a measure of the ability of overseas health professionals to cope with the English language demands of their particular medical specialisation. The inferences to be drawn from performance on the test were therefore related to their language competence, no more. In this respect, the OET appears to have been a successful test. If it were to become a ‘true’ performance measure (i.e. a measure of the test taker’s ability to perform the particular medical duties under scrutiny) then clearly a different approach to the evaluation of the performance would be needed.

As has been suggested elsewhere (e.g. Davies, 2001), there is little evidence to support the view that language for specific purposes is radically different to the language of everyday life (though as we can see in Douglas’s suggested theoretical rationale for LSP testing, there is certainly evidence that there are quantifiable differences between ‘general purpose’ and ‘specific purpose’ language). It would therefore seem reasonable to use scales developed for ‘general purpose’ use for the evaluation of ‘specific purpose’ language, with the provision of a different expectancy. This would result, for example, in a rater encountering a criterion designed to focus on vocabulary and then applying that criterion within the context of the specific purpose being tested.

The logic of the argument proposed by Douglas (and by Jacoby & McNamara) would result in an inability to test language without also testing the test taker’s ability to perform within the context of the specific area being tested. There is a danger here of slipping into what McNamara (1996) describes as a ‘strong’ view of performance tests, where the task is the focus of the test, with language seen only as one of a number of abilities needed to successfully complete or execute that task. In this case, the problem of inseparability of language and non-language factors influencing the evaluation of performance by raters is magnified (see Elder, 2001 for an extended discussion of this area). Basically, the problem arises from potential contradictions in the requirements of the different aspects of the ‘strong’ performance test. In the ‘weak’ view of performance testing (advocated by McNamara, 1996) the focus is solely on the language of the interaction. Here, there is a clear differentiation made between the test takers’ language ability (as measured within the specific

purpose domain) and their ability to perform a task related to that same domain.

In terms of business language testing, and, in fact, of tests of any other specific purpose context, it would be extremely difficult for a test developer to create a valid instrument, where the construct definition fails to clearly distinguish language and specific purpose abilities.

### Re-defining business language tests

It is not helpful to take the view that tests can only be seen as being ‘specific purpose’ if they are very narrowly focused on a particular ‘purpose’ area and are representative of what McNamara (1996, 1997) sees as the ‘strong’ view of performance testing – where the language and the ‘purpose’ are both considered as part of the construct. Instead, there are two alternative views of ‘specific purpose’ tests that offer a not incompatible expansion to the definition of SP tests offered by Douglas (2000: 19).

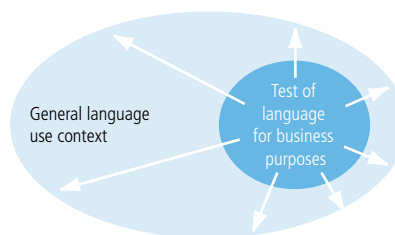
**View 1:** as all tests are in some way ‘specific’, it is best to think of language tests as being placed somewhere on a continuum of specificity, from the broad general purpose test (such as CPE) to the highly specified test, such as the Japanese Language Test for Tour Guides.



**View 2:** very highly specific tests tend to be very poor in terms of generalizability, while the opposite can be said of non-specific tests. There is not a binary choice in operation here, and if we accept that tests can be developed along a specificity continuum, then it logically follows that a test which is placed somewhere other than the extremes of the continuum will have the potential to be either more or less generalizable.

We could conceive of a test that is specific only in that it is placed within the context of an employment/career area (for example ‘business’), and that will be generalizable to the broader ‘general language use’ context (see Figure 1).

Figure 1. Potential generalizability from less specifically defined test to the broader context of general language use



This is the case, for example, with the UCLES business language tests which, as the following articles will show, are specific purpose in as much as they are located within a business employment context and so reflect features of that context.

## References and further reading

- Berry, V (1996): *Ethical considerations when assessing oral proficiency in pairs*, paper presented at the Language Testing Research Colloquium
- Berry, V (1997): *Gender and personality as factors of interlocutor variability in oral performance tests*, paper presented at the Language Testing Research Colloquium
- Brown, A (1998): *Interviewer style and candidate performance in the IELTS oral interview*, paper presented at the Language Testing Research Colloquium
- Davies, A (2001): The logic of testing languages for specific purposes, *Language Testing*, 18/2, 133-147
- Douglas, D (2000): *Assessing Language for Specific Purposes*, Cambridge: Cambridge University Press
- Douglas, D (2001): Language for specific purposes assessment criteria: where do they come from? *Language Testing*, 18/2, 171-185
- Elder, C (2001): Assessing the language proficiency of teachers: are there any border controls? *Language Testing*, 18/2, 149-170
- Jacoby, S (1998): *Science as performance: socializing scientific discourse through the conference talk rehearsal*, Unpublished Doctoral Dissertation. UCLA
- Jacoby, S and McNamara, T F (1999): Locating competence, *English for Specific Purposes*, 18/3, 213-241
- Labov, W (1963): The social motivation of sound change, *Word*, 19, 273-307
- McNamara, T F (1996): *Measuring Second Language Performance*, London: Longman
- O'Sullivan, B (1995): *Oral Language Testing: Does the Age of the Interlocutor make a Difference?* Unpublished MA Dissertation. University of Reading
- O'Sullivan, B (2000): *Towards a model of performance in oral language testing*. Unpublished PhD Thesis. University of Reading
- O'Sullivan, B (2002): Learner acquaintanceship and oral proficiency test pair-task performance, *Language Testing*, 19/3, 277-295
- Porter, D (1991a): Affective Factors in Language Testing. In Alderson, J. C. and B. North (Eds.) *Language Testing in the 1990s*, London: Macmillan (Modern English Publications in association with The British Council), 32-40
- Porter, D (1991b): Affective Factors in the Assessment of Oral Interaction: Gender and Status. In Sarinee Arnivan (ed) *Current Developments in Language Testing*. Singapore: SEAMEO Regional Language Centre. Anthology Series 25: 92-102
- Porter, D (1994): LSP testing – luxury or necessity? In Khoo, R. (ed). *LSP Problems & Prospects*, Singapore: SEAMEO RELC, 194-201
- Porter, D and Shen Shuhong (1991): Gender, Status and Style in the Interview, *The Dolphin 21*, Aarhus University Press, 117-128
- Swales, J (1990): *Genre analysis: English in academic and research settings*, Cambridge: Cambridge University Press
- Tarone, E (1985): Variability in Interlanguage use: A study of style shifting in morphology and syntax, *Language Learning*, 35/3, 373-404
- Tarone, E (1988): *Variation in Interlanguage*, London: Edward Arnold
- Tarone, E (1998): Research on interlanguage variation: Implications for language testing. In L. F. Bachman and A. D. Cohen (Eds.) *Interfaces Between Second Language Acquisition and Language Testing*. Cambridge: Cambridge University Press, 71-89
- Tarone, E and Liu, G Q (1995): Situational context, variation, and second language acquisition theory. In G. Cook and B. Seidlhofer (Eds.) *Principle and Practice in Applied Linguistics*, Oxford: Oxford University Press. 107-124
- Teasdale, A (1994): Authenticity, validity, and task design for tests of well defined LSP domains. In Khoo, R. (ed). *LSP Problems & Prospects*. Singapore: SEAMEO RELC, 230-242

# Revising the Business English Certificates (BEC) speaking tests

DAVID BOOTH, SPECIALISED EXAMS GROUP

## Introduction

This article reports on recent work to revise the Business English Certificate (BEC) speaking tests within the context of the overall revision of the BEC suite of examinations.

Previous articles in *Research Notes* have dealt in detail with issues relating to test development and revision projects (Nick Saville – *Research Notes 4*, Nick Saville and Barry O'Sullivan – *Research Notes 3*, and Lynda Taylor – *Research Notes 3, 4, 5 and 6*). In *Research Notes 4* Nick Saville summarised the process whereby UCLES EFL tests are routinely revised following a cyclical, iterative model of development. BEC was revised in line with this model.

The Business English Certificates were introduced between 1993 and 1996. The tests were originally developed for the Asia-Pacific

region, in particular China; as a result, the content of the tests related to the local conditions in which the test was taken and the topics and tasks were made accessible to the test-taker population. In 1998 the tests were made more widely available to meet the needs of a worldwide candidature seeking qualifications in English within a business context. In line with UCLES' policy of reviewing its tests at regular intervals, a full review of BEC was begun in 1999.

## Groups involved in the revision process

A number of groups were involved in the review and revision process, and they played a crucial role in the development of the revised specifications for the speaking tests. Their roles and responsibilities are detailed below:

The **Management Steering Group** – made up of UCLES EFL senior management. The role of this group was to define parameters, initiate research and development, make judgements, and in the final stages of the revision ratify revised specifications.

An **Internal Working Group** – made up of UCLES EFL specialist staff including research/validation staff. The role of this group was to co-ordinate external groups, act on recommendations from the steering group, trial revised specifications and report back to the steering group.

**Consultants working groups** – made up of UCLES EFL consultants, specialist staff and research/validation staff. The role of these groups was to devise revised specifications for each component. The groups met as required to develop the specifications.

## Issues for internal consideration

Over recent years, UCLES EFL has been seeking to harmonise the principles and practice which apply to the range of face-to-face speaking tests it offers. At the outset of the BEC revision process, therefore, there were two key internal issues relating to the revision of the speaking tests.

### 1 Test level

The original BEC 1 had been designed to span 2 proficiency levels – Common European Framework (CEF) levels A2 and B1. Since there was a limited amount of ‘business’ language which could be tested at A2, the BEC 1 speaking test had been developed to include aspects of general English as well as English in a workplace context. As part of the review process, it was decided that it would be more appropriate to develop the BEC Preliminary Test at CEF B1 only. This would help to align it more closely with the Council of Europe Framework and ALTE proficiency levels and, as a result, help to make the speaking test more business-focussed.

### 2 Score reporting

For historical reasons, the grade for the current BEC speaking test was awarded separately from the rest of the test score and did not contribute to the overall score or grade of the candidate. It was decided that the speaking mark should be incorporated into the overall mark for the test and that the speaking test should be given the same weight as the other components, as is the case for the other Cambridge examinations. This, it was hoped, would have a positive effect on classroom practice.

## Consultation with external stakeholders

The development of BEC into a global examination had generated a certain amount of formal and informal feedback on the test construct and testing systems. The level of feedback was not sufficient, however, to form a basis for decision-making so a wider variety of opinions and views needed to be collected. In the early stages of the revision process, an extensive programme of consultation was initiated to ensure that any changes to BEC would be in line with the expectations of stakeholders. In addition to

consultation via the consultants working groups referred to above, two questionnaires were sent out, one to centres and the other to a small number of key stakeholders.

### 1 BEC centre questionnaire

The first questionnaire was designed to be sent to BEC centres and was largely answered by teachers. It focussed on evaluating general levels of satisfaction with the current test and it also gave respondents an opportunity to make suggestions on how the test could be improved. The questionnaire was sent to almost 400 centres of which 70 responded.

Overall, the results suggested that the test was well liked. The satisfaction rates for the three tests are shown in Table 1 below.

#### Question: How satisfied are you generally with BEC?

Table 1: General levels of satisfaction for BEC

Test	Very Satisfied or Satisfied	Dissatisfied	Very dissatisfied
BEC 1	90%	10%	0%
BEC 2	85.3%	8.8%	5.9%
BEC 3	81.8%	18.2%	0%

The levels of satisfaction with the speaking tests were much lower, however, as can be seen in Table 2 below.

#### Question: How satisfied are you generally with the BEC speaking test?

Table 2: Levels of satisfaction for BEC speaking tests

Test	Very Satisfied or Satisfied	Dissatisfied	Very dissatisfied
BEC 1	55.1%	34.5%	10.3%
BEC 2	58%	25.8%	16.8%
BEC 3	70.6%	23.5%	5.9%

The questionnaires allowed respondents to add comments to their evaluations. Selected comments are shown below to illustrate some of the main concerns.

#### Comments relating to the BEC 1 Speaking Test :

Country	Comment
France	It seems a little too general.
Argentina	It is too limited and it does not match the demands in reading/writing abilities. There should be another phase showing a better command of Business vocabulary.
Germany	There is no scope for candidates to expand; both the “script” and candidate output is likely to be repetitive and mechanical.
Spain	Needs to be less ‘general’ in content and more ‘working world’ specific.
Bangladesh	Not challenging enough – not related to business.
Italy	Can it not be more ‘businessy’?

### Comments relating to the BEC 2 speaking test:

Country	Comment
UK	Booking a hotel is irrelevant; candidates should have a chance to express business knowledge i.e.: stock markets/international trade.
Argentina	Once again the situations do not always reflect the communicative needs of someone working for a company. I find the follow-up planned discussion quite helpful however.
Italy	Not challenging enough – rigid framework does not allow candidates to express themselves fully.
Spain	We do not like the 2 way collaborative task – it is artificial, does not elicit natural language and it is easy to just read responses; we feel it definitely needs re-thinking – the students always seem confused by this activity.
Czech Republic	I have heard that it is a bit easy & doesn't really match or measure the rest of the test.

### Comments relating to the BEC 3 speaking test:

Country	Comment
Spain	Enjoyed the freedom given to candidates to talk freely. Choice of topic was fine.
Argentina	However the first part is still too general (giving personal information)
Spain	Maybe some of subject matter is difficult for pre-work experience students.
Czech Republic	Too easy, not enough points.

## 2. Key stakeholders questionnaire

The second questionnaire was aimed at people with a detailed knowledge of the BEC examination who were also specialists in the field of testing. The key consultant group consisted of:

- external testing specialist consultants involved in developing the paper (chairs and principal examiners);
- Senior Team Leaders (responsible for the training and evaluation of Oral Examiners);
- UCLES business development managers;
- administrators in large BEC centres.

The purpose of this questionnaire was to obtain more focussed information on particular features of the BEC tests, including suggestions for development of the test. The stakeholders were asked to respond to statements about the test on a scale of 1 to 5, where 5 indicated strong agreement and 1 strong disagreement. Overall, there were fifteen features of the test rated above 4 and therefore rated as very satisfactory. Seven features of the test attracted an average rating below 3 (i.e. ratings expressing a measure of dissatisfaction with the existing test); and five of these related directly to the speaking tests. On average, respondents

tended to *disagree* or *disagree strongly* with the following statements about the speaking test:

*It is useful to have the speaking skill reported separately from reading, writing and listening*

*It is useful that BEC 1 covers both Cambridge levels 1 and 2*

**BEC 1 Speaking:** *The activities in part 2 are an effective way of testing spoken language*

**BEC 1:** *The speaking test gives candidates sufficient opportunity to show their full language ability*

**BEC 2 Speaking:** *The speaking test gives candidates sufficient opportunity to show their full language ability*

## Conclusions from the questionnaire surveys

The consultation exercise confirmed that there were some concerns among external stakeholders over both the level of BEC 1 and the separate grade given for the speaking paper. These issues had already been addressed by steering group recommendations. Further information gained from the questionnaires supported earlier informal evidence, in particular:

- the inadequate business focus/content of the speaking test, particularly at BEC 1;
- the inadequate amount/quality of language produced by the information gap activity in BEC 1 and BEC 2.

## Redefining the specifications

The issues raised through the questionnaires and working groups were addressed during the development stage of the revision project and the key outcomes are discussed below:

### Ensure the tests are clearly business focussed

This was particularly an issue at Preliminary (ALTE level 2) and Vantage level (ALTE level 3). For both tests the Interlocutor frames for Part 1 were reformulated to introduce more business language earlier. The preliminary frames may still include one general question to introduce a topic but all the questions in the frames now focus on business situations. At all levels the 'long turn' (Part 2) and discussion topics (Part 3) are clearly business focussed.

### Include a 'long turn' in which candidates have the opportunity to produce longer utterances

As can be seen from the feedback, it was felt that the 'information gap' exercise in BEC 1 and 2 did not produce enough extended discourse for the purpose of assessing discourse features. The short presentation in BEC 3, however, did seem to provide this. The revision team felt that, with adequate support, the short talk would be the most appropriate way of eliciting extended discourse at all three levels. It was also felt that the task type would impact positively on the classroom where short talks and presentations are often included in business-focussed courses.

Trialling confirmed that the short presentation tasks which were

developed and the timings which were allocated were appropriate for candidates at Preliminary and Vantage levels.

### Ensure a discussion topic is included at each level to allow a wider range of interactive linguistic features to be assessed

The increased time allowed for the test made it easier to include a discussion task at each level. A number of options were considered, including the use of pictures to stimulate discussion. The preferred model was discussion generated by a written prompt. For BEC Vantage a similar task to Higher was developed. For the Preliminary level, however, it was felt that some tasks would be better suited by a format which allowed the use of illustrations.

In addition to matters of test content and format, other issues addressed included:

### Clarifying the appropriate assessment criteria to use when assessing language in a business context.

When developing EFL examinations, UCLES takes account of current theories of language, language learning and good practice in assessment. Much of the relevant theory with regard to a model of language can be found in the Council of Europe's Common European Framework of Reference (2001). Key to an understanding of the construct underlying BEC is the recognition of Business English as a 'context of use' or domain of language as described in the Framework documentation. An internal document was drawn up by Dr Lynda Taylor regarding the relationship between the

construct of the BEC speaking tests and the main suite speaking tests. She concluded:

*'The underlying construct of spoken language ability is therefore common to both general English language proficiency tests (KET, PET, FCE, CAE and CPE) and the BEC suite. The difference between the two suites lies in the various content features of the tests: choice of vocabulary, themes (or topics), purposes, text types, functions, and communicative tasks and situations presented in the tasks.'*

(See also the article on page 2 in this issue by Barry O'Sullivan.)

### Ensuring the production quality values of the test materials meet the expectations of candidates and clients.

Given the expectations of business candidates it was felt important that the speaking test materials needed to be well presented. With this in mind the speaking material formats were redesigned.

## Conclusions

This article has highlighted some key issues and outcomes of the revision of the BEC speaking tests. A full account of the revision process will be published in an upcoming volume of the *Studies in Language Testing* series.

## Reference

Council of Europe (2001): *Modern Languages: Learning, Teaching, Assessment. A Common European Framework of reference.*

# Revising the BULATS Standard Test

ED HACKETT, SPECIALISED EXAMS GROUP

## A brief history of BULATS

BULATS (Business Language Testing Service) is a multi-band test of language proficiency in the workplace for in-company use. It is available in four languages (English, French, German and Spanish) and has four test components: the Standard Test (a paper and pencil based test of Listening, Reading and Language Knowledge), the Computer Test (a computer adaptive version of the Standard Test), the Speaking Test, and the Writing Test. Components can be taken individually or together, depending on company needs, and measure language proficiency on the ALTE scale 0 to 5 (beginner to upper advanced), which corresponds to the Council of Europe Framework A1 to C2. BULATS is sold through a world-wide network of Agents and is available on demand. It is marked locally, which means results can be produced within days, or instantaneously in the case of the Computer test.

The BULATS Standard Test was first launched in 1997 in a small number of countries, but the network has since grown to over 50 Agents in 30 countries. The Speaking and Writing tests have been available since 1998 and the Computer test was launched in 2000.

## Test design and construct

The construct of BULATS is similar to that of its sister test BEC (a certificated test available to individual candidates) in that it is a test of language in the workplace, rather than a test of specific business language. In its original design, the Standard Test comprised three sections: Listening, Reading, and Grammar and Vocabulary. The test had 90 items, lasted 90 minutes and reported an overall score on a scale of 0 to 100, which translated to an ALTE band (see Figure 1).

Figure 1 : BULATS Scores and ALTE Bands

BULATS Score	ALTE Band	Level
90 to 100	5	Upper Advanced
75 to 89	4	Advanced
60 to 74	3	Upper Intermediate
40 to 59	2	Lower Intermediate
20 to 39	1	Elementary
0 to 19	0	Beginner

In response to demand from clients, section scores were also given on a converted scale out of 50. The aim of the section scores was to give clients information on differential performance on the three skills, but was not originally intended to be used as an indicator of performance in test-retest situations. The main function of the Standard Test was to give companies or training organisations a quick yet reliable indication of general language proficiency. For more informed information on full language competency, clients are recommended to use the Speaking and Writing components in addition to the Standard or Computer Test.

## The need for change

As with any test, candidature and usage change over time, and during the past three years a more detailed picture has emerged of the BULATS population and the needs of our clients. Between January 2000 and January 2001, a major calibration exercise took place involving post test analysis of Answer sheets of more than 6000 candidates in over 10 countries. Special calibration forms of the test, combining sections of past and new papers, were also used on live test populations to verify the equating of live versions. Key Agents were given questionnaires regarding test usage and needs.

The findings of the post test analysis and questionnaire revealed that the test was working well overall and that there was a high degree of contentment from clients. There was, however, a shift in usage of the test from general proficiency testing to progress testing. Whilst this did not present a problem for overall scores, it was felt that the test design could be improved to increase the reliability of the section scores. The overall reliability of the test was very high (Standard Test versions report overall reliability alphas of between 0.95 and 0.96). The section scores, whilst respectable (with alphas of between 0.85 and 0.92), were less reliable than overall scores when standard error is taken into account. As mentioned above, the section scores were originally provided to allow comparison between the different skills tested and were not meant to provide measurement of progress. The overall test has 90 items, which provides sufficiently broad band widths to allow for Standard Error of Measurement (SEM), but as the section scores have fewer items (between 25 and 35) reliability is less accurate. The problem was more pronounced in the Reading section, which only had 25 items, leading to a bunching of abilities in bands 3 to 5. In addition to improving reliability in the section scores to take account of current test usage, it was also felt that the test design could be improved to cater for the wide range of abilities of the candidates.

## Options for change and revision constraints

Drawing on the experience of the BEC revision process, it was felt that the test could be revised in a relatively short space of time. The four key areas to consider in the revision process were Validity, Reliability, Impact and Practicality. The key factor here was reliability, yet improvements in reliability would impact on other areas. BULATS has high face validity with both clients and candidates due to its variety of communicative tasks. In addition to discrete point multiple choice tasks, there are integrative multiple choice, matching and productive tasks. Clients also like the fact that they can use BULATS anytime and anywhere, so lengthening the test dramatically, or changing the marking system would have consequences for both the impact and practicality of the test. Feedback from Agents had stressed a high degree of contentment with the existing format. So, whilst improving section reliability was the main aim of the revision, there were a number of constraints which had to be taken into account. These were:

- the test should not be lengthened beyond a 2 hour limit;
- the test should report at least two section scores, with a discrete listening section;
- the test should comprise grouped tasks as well as discrete items;
- the test should be clerically marked;
- the revision process should not take more than 18 months.

## Test revision

Analysis was done using the Abils and Best Test programs (developed by Dr Neil Jones for UCLES) and it was decided that a minimum of 50 items was needed to provide section scores with the desired improvement in reliability. A greater spread of item difficulties gave suitable band widths at all levels, providing better discrimination in bands 1 and 5. This raised a problem for the format of the test, as retaining three sections with 50 items and the current task types, would produce a test lasting nearly 3 hours. The only way to produce a test with 150 items in under two hours would be to change the majority of items to discrete point task types, and this would impact negatively on both construct and face validity. The decision was therefore made to combine the Reading, and Grammar and Vocabulary sections (60 items) and expand the Listening section to 50 items.

## Changes to the Listening section

10 graphic and text-based discrete items were added to the start of the test, to give better discrimination at lower levels, and the one 8-item, multiple choice long text was replaced by three 6-item tasks, allowing for more items targeted at bands 3 to 5. This then presented a problem with timing, as it was felt that a Listening section of 60 minutes was too long, so the decision was made to change some of the tasks to 'once only' listenings. There has been much debate over the use of listen-once or listen-twice tasks and there are cogent arguments on both sides. In real life we sometimes



have the opportunity to ask for clarification of points we have not fully understood from the first hearing, and announcements or news headlines are sometimes repeated. However, the vast majority of what we listen to is heard once only, so it can be argued that... 'in terms of both situational and interactional authenticity of the language, playing the text just once seems to be the obvious thing to do.' (Buck, 2001). Past UCLES trialling into the impact of listen-once versus listen-twice suggested that both forms spread out candidates' abilities in the same way, but listen-once increased the difficulty of the majority of tasks slightly. It was felt that there were strong arguments for retaining the listen-twice format for some tasks, but that other task types could be converted to the listen-once format with minor changes. It was decided that form-filling tasks, where key information is deliberately stressed or spelled out, and multiple matching tasks, which test the gist of a text, were best suited to once only listening, and these tasks were trialled. The effect on item difficulty proved to be minimal across both task types with an average increase in difficulty of only 0.08 logits. Multiple choice discrete and grouped tasks were kept as listen-twice. The discrete items come at the beginning of the test and are aimed at lower ability candidates. It was felt that double listening helped compensate for the lack of scene setting that would be available in real life, and allowed candidates to 'tune in'. The content in short texts can easily be missed by low level candidates, and this can have a demoralising effect if heard only once. The grouped multiple choice tasks (6 items) involve a heavy reading load and the second listening allows time to re-read the options and catch up on parts of the text missed while reading the questions in the first listening. Further research into the effect of listen-once versus listen-twice is ongoing. The effects of the above changes meant that the 50 items could be fitted into 50 minutes of listening time.

### Changes to the combined Reading and Grammar and Vocabulary section

A number of changes to both the format and task types were made in the newly-titled Reading and Language Knowledge section. Reading and Language Knowledge tasks were alternated to avoid the negative effects of lack of time and fatigue on the grammar and vocabulary tasks, which had originally been at the end of the paper. Tasks were also more finely graded from easy to difficult to allow candidates to find their natural level. The section was divided into two parts, with items aimed at bands 0 to 2 in Part One, and items aimed at bands 3 to 5 in Part Two. Trialling of the revised format revealed that all candidates in bands 0 to 2 had time to complete the first part, and 97% of candidates in bands 3 to 5 completed the whole paper in the time allowed.

The Reading section had proved the most problematic part of the Standard Test. In addition to the small number of items (25), there were not enough items discriminating well at high levels. The discrete graphic items and the matching sentences to paragraphs task were well-suited to low level candidates. However, the 8-item multi-task long text, aimed at candidates in bands 3 to 5, did not discriminate sufficiently well at these levels. It was decided to

replace this task with two 6-item multiple choice texts (one at higher and one at lower level). Such task types have proved better suited to discriminating between narrower ability bands. Evidence from BEC3 (now BEC Higher) and CPE suggests that these tasks discriminate well at high levels.

The 15-item multiple choice cloze had proved difficult for lower level candidates, who found the length of the text daunting. This meant that some candidates gave up on the passage, even though there were items within the task at their level. This was replaced by a 5-item multiple choice cloze. The 10-item Open Cloze was dropped in favour of two 5-item clozes, one higher and one lower level. Discrete lexico-grammar-based multiple choice tasks were added to both parts of the section to provide a more gradual increase in item difficulty.

Figure 2: Comparison of existing and revised formats

Existing Format		Revised Format	
<i>Section 1 Listening</i>		<i>Section 1 Listening</i>	
Part One	2 x 5-item Multiple Matching	Part One	10 x discrete text and graphic MC
Part Two	3 x 4-item Form Filling	Part Two	3 x 4-item Form Filling
Part Three	1 x 8-item long text MC	Part Three	2 x 5-item Multiple Matching
		Part Four	3 x 6-item long text MC
<i>Section 2 Reading</i>		<i>Section 2 Reading &amp; Language Knowledge</i>	
Part One	10 x discrete graphic texts	Part 1.1	7 x discrete graphic texts
Part Two	1 x 7-item matching	Part 1.2	6 x discrete MC Lexico-grammar
Part Three	1 x 8-item multi-task	Part 1.3	1 x 6-item MC Reading text
		Part 1.4	1 x 5-item Open Cloze
		Part 2.1	1 x 7-item matching
		Part 2.2	1 x 5-item MC Cloze Correction
		Part 2.3	1 x 5-item Open Cloze
		Part 2.4	6 x discrete MC Lexico grammar
		Part 2.5	1 x 6-item MC Reading text
		Part 2.6	1 x 7-item Error Correction
<i>Section 3 Grammar &amp; Vocabulary</i>			
Part One	1 x 15-item MC Cloze		
Part Two	1 x 10-item Open Cloze		
Part Three	1 x 10-item Error		
Items	90	Items	110

The revised format of the test was extensively trialled from November 2001 to February 2002, and feedback from the Agents and candidates involved has been encouraging. The revised format is due to be launched at the end of May 2002. Further details about BULATS can be found on the website: [www.bulats.org](http://www.bulats.org)

### Reference

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

# Developing wordlists for BEC

FIONA BALL, RESEARCH AND VALIDATION GROUP

## Introduction

This article describes the ongoing development and validation of vocabulary lists for UCLES' Business English Certificates, focusing on the BEC Preliminary level.

Wordlists exist in many forms and serve a variety of purposes. In the context of examining English as a Foreign Language, wordlists have two main applications. Firstly, they indicate acceptable vocabulary and structures to be used by item-writers in developing examination materials. The second application of wordlists is more research focussed, that is to study the words and structures produced by candidates in live examinations. This has obvious applications in teaching and publishing since knowledge of productive vocabulary can be used to inform teaching strategies and coursebooks that prepare candidates for an examination at a particular level.

There are many questions that arise from attempts to define the level and amount of vocabulary necessary to succeed at a particular examination. UCLES EFL is therefore currently investigating:

- What is the nature of the difference between productive and receptive vocabulary?
- How do individual candidates differ in their productive vocabulary?
- How can business vocabulary be distinguished from general vocabulary?

This article focuses on how we are investigating the extent and complexity of vocabulary related to the BEC suite of examinations. The most recent update of the BEC Preliminary wordlist and insights gained from studying the written production of BEC candidates are described.

## The written component of the BEC Preliminary examination

The BEC examinations are 'aimed primarily at individual learners who wish to obtain a business-related English language qualification' (*BEC handbook*, p.2). BEC Preliminary is aligned with the ALTE/Cambridge level 2 so tests the same level of proficiency in English as the Preliminary English Test (PET), a general English examination. Topic areas covered by BEC include the office, general business environment and routine, travel and meetings, and such topics naturally influence the vocabulary used in examination materials and expected of the candidates. At any BEC level the item-writers aim to provide authentic sounding materials that are accessible to the whole candidature.

The writing component of the BEC Preliminary examination requires candidates to write a short internal company communication, such as a memo or email, followed by a longer piece of business correspondence with an external contact, such as a letter or fax. These business-related tasks include a written prompt and bulleted points show candidates what is to be included in their answer. Item-writers consult the BEC Preliminary wordlist to check what vocabulary they can include in the examination materials.

## The BEC Preliminary item-writer wordlist

The BEC Preliminary item-writer wordlist aims to cover business-related vocabulary relevant to this level of Business English. Item-writers refer to this list when producing materials for the examination. They also have access to the PET list as this includes words and structures at the same level, albeit in the scope of general English rather than Business English. The wordlist includes parts of speech together with examples that highlight one particular sense of a word, for example:

<i>address (n) (v)</i>	<i>To address a conference</i>
<i>advance (adj)</i>	<i>in advance/advance booking</i>
<i>air (n)</i>	<i>by air</i>

There is also a list of suffixes and prefixes and a list of word groups within the list that further specify the extent of vocabulary that item-writers can draw on.

One of the issues we face in developing any item-writer wordlist is ensuring that a list is equally appropriate for developing speaking, listening, reading and writing components of a specific examination. For the BEC Preliminary wordlist, we seek to include *current* business usage wherever possible; this can change rapidly, as shown by the recent growth of email correspondence and use of associated lexis in the last decade. We maintain and develop the BEC Preliminary wordlist by adding and removing words and affixes on a regular basis. Words are suggested for inclusion in, or exclusion from, the list each year. This procedure is informed by corpus data and detailed discussion, as described below.

## Corpus-based approaches to wordlist development

Our development and validation of item-writer wordlists draw on corpus evidence in addition to the experience and knowledge of the chairs of the item-writing teams for BEC. A range of corpora (electronic databases) that represent receptive and productive language, business and general English, and learner and native

speaker data are explored. The methodology involves exploring corpus examples and frequency measures and considering comparative and raw frequencies for each word under consideration. A new data source available for the 2002 revision of the BEC Preliminary wordlist was a list of words derived from all of the BEC writing in the Cambridge Learner Corpus (CLC), a large corpus of candidates' writing scripts from Main Suite and BEC examinations. The BEC corpus-derived wordlist is the focus of another current research project and will be described in a future article.

The BEC Preliminary wordlist is developed in several stages. Suggested additions to the wordlist are collated over a six month period and the raw frequency of these words in a range of learner and native speaker English corpora is obtained. Next, the list of suggested words is compared against other item-writer wordlists (for KET and PET) and against CLC-derived lists that illustrate the written production of a large number of candidates taking KET, PET, FCE and BEC examinations. Each word is tested against various frequency criteria and the resulting data guide the final discussion in which words are added to or kept out of the wordlist.

In the 2002 revision, approximately fifty words were considered for inclusion in the BEC Preliminary wordlist. These were searched for in various corpora in order to determine their frequency and behaviour in various types and levels of English. The four corpora investigated were:

- British National Corpus (BNC) (100 million words of written and spoken native speaker English, including 4 million business oriented words);
- Cambridge Learner Corpus (CLC) (14 million words of written learner English at five levels for general and business English);
- Business Texts Corpus (120,000 words of U.S. and U.K. business articles from 1998);
- Trial Web Corpus (11,000 words of contemporary business English from the Internet October 2001).

Each corpus provided a unique perspective. The CLC provided different senses of each word and examples from learner texts, thereby showing whether the suggested words were already being used productively by BEC candidates. The BNC indicated how frequent each word was in native speaker data, in both productive and receptive contexts, although this was based on a mixture of older and more contemporary texts. The two business corpora focussed on the use of the suggested words in business texts, and also provided a comparison of usage between 1998 and 2001.

Once the list of suggested words for inclusion was obtained, the CLC was explored for evidence of these words being used productively by candidates. At this stage the suggested words were checked against the BEC wordlist derived from the CLC and this corpus was searched to provide contextualised examples. A ranked table of frequency in the CLC was produced to provide a measure of the raw and normalised frequency of the suggested words across different levels of business and general English in UCLES' examinations. The term 'normalised' refers to a weighted frequency

measure that allows for easy comparison between two sets of data of different sizes. In the CLC, for example, there are half a million words of BEC whilst there are over 10 million words of Main Suite. This imbalance is significantly reduced when the BEC and Main Suite figures for each suggested word are re-calculated as frequencies per 10 million words. Table 1 shows the normalised frequency for several words in the learner data that were considered for the BEC Preliminary wordlist in the last revision.

Table 1 : Normalised frequencies in the CLC

Word	Normalised Frequency		Total
	General	Business	
globe	185	31262	31447
culture	7259	1045	9305
level off	0	2810	2810

This table shows that *globe* and *level off* are much more frequent in business than general learner English. The opposite is shown by *culture* which occurs more frequently in Main Suite than BEC writing. The fact that *level off* does not appear at all in the Main Suite data suggests that this might be part of a *core business vocabulary*, a concept which we are seeking to explore further.

The suggested words were then investigated in native speaker data using frequency lists based on the British National Corpus (see Adam Kilgarrif's website). Table 2 shows the five most frequent suggested words in the BNC.

Table 2 : Raw frequencies in the BNC

Word	Raw Frequency
maintain	11881
determine	11551
sector	10937
culture	10196
access	10099

These raw frequencies were converted into a normalised frequency count so that these figures could be compared with those for the Learner Corpus data in Table 1. The next stage was to consider the frequency of each suggested word in business English texts within the trial web corpus developed for this purpose and the existing business texts corpus which contains slightly older data. Table 3 shows the raw frequency of some of the suggested words in the trial web corpus. It is important to note that these figures are lower than those in the previous two tables because this corpus is much smaller; nevertheless, it was the most relevant to this project as it reveals current business usage which the other corpora cannot provide.

Table 3: Raw frequencies in contemporary Business English

Word	Raw Frequency
service	28
sector	7
creative	4
e-commerce	2
extend	2

A range of evidence was therefore produced from the three analyses presented above. Firstly, a ranked list of the normalised frequency of all of the suggested words was produced. A table was then produced showing the ranked raw frequency of each suggested word in each of the four corpora in order to assess their relative frequency in the data as a whole. The top ten words out of the fifty under investigation were singled out for further analysis in each corpus list. Where this excluded many words from further discussion, the top 20 words in each CLC-derived list were then considered as these were words that had already been used productively by a number of candidates in live examinations. A list of words to be further considered for inclusion in the BEC Preliminary Wordlist was therefore derived based on:

- Words with a frequency of >500 per 10 million in all four corpora;
- Words that occur in two or more top ten lists;
- Words that occur in the top 20 words of Main Suite exams;
- Words that occur in the top 20 words of BEC exams.

This quantitative evidence was discussed by a panel of BEC chairs and subject officers in order to reach a final decision on which words should be included in the item-writer wordlist. The result of this meeting was that around thirty new words were added to the BEC Preliminary item-writer wordlist, based on the frequency and corpus evidence described above.

Despite the success of this procedure in using corpus-evidence and experience together, there are some limitations to the methodology adopted here. The fact that there is no large contemporary corpus of business English means that a sample will have to be collected from suitable sources on an annual basis in order to illustrate current usage. Secondly, the Learner Corpus is not tagged for parts of speech; this means that examples have to be consulted in order to determine which sense or part of speech a candidate has used. With learner data, there is also the chance that a word has been used incorrectly or inappropriately, but this can also be checked by using corpus evidence.

Methodologically, normalised frequency measures are less reliable with small corpora and although we can be certain of the authenticity and representativeness of our own corpora, the other corpora used were built with other aims in mind. A final disadvantage of using learner data is that the task-effect is thought

to have a strong influence on the productive vocabulary of candidates; this will be investigated in a future project by comparing the language of individual candidates with that provided by the task, using WordSmith Tools – a text analysis software package (see website).

## Future research into wordlists

In addition to regularly updating the BEC Preliminary and KET and PET item-writer wordlists, UCLES is currently investigating the differences between productive and receptive vocabulary. This project uses the corpus-derived wordlists from the CLC for Main Suite and BEC examinations. The CLC-derived BEC wordlist initially contained more than 20,000 different types (based on half a million tokens) and required significant sorting. A range of analyses are currently being undertaken on this list which aim to give a truer picture of BEC productive vocabulary. The first stage is to remove all names, non-English and unrecognisable words from the wordlist to reveal the core and most frequent vocabulary requiring further investigation. The second stage will involve dividing the remaining headwords according to core and business English vocabulary to enable comparison between general and business English. Furthermore, UCLES aims to identify a subset of words that represent different levels of Business English. The results of this research will be reported in future issues of *Research Notes*.

It is hoped that the results of this research will help UCLES to define business English more rigorously than has been done to date. Whilst vocabulary knowledge alone does not represent all that is needed by a candidate to communicate effectively in an English-speaking business context, it is nevertheless a key aspect of the difference between general and business English.

## Conclusion

Several key questions now face us in relation to vocabulary:

- What constitutes the business component of BEC writing output?
- How can a core business vocabulary be described and defined?
- How is this core vocabulary distributed across the three levels of BEC?
- To what extent does the task influence BEC productive vocabulary?

We are currently working to identify and investigate both productive and receptive vocabulary across different levels and for different types of English; we are also investigating the notion of a significant frequency measure for vocabulary. For productive data we will continue to use the Cambridge Learner Corpus whilst receptive data will be obtained from UCLES' computerised bank of items (see *Research Notes 1* and *2*) and surveying appropriate course-books. Once we have determined a measure of significant frequency, we may be able to specify more clearly the range of words that an average candidate at any level should be able to produce or recognise.

## References and further reading

Adam Kilgarriff's website: BNC data and frequency lists  
<http://www.itri.bton.ac.uk/people/index.html>

BEC handbook (UCLES 2001) downloadable from:  
<http://www.cambridge-efl.org.uk/support/downloads/bus.cfm>

Business Texts Corpus:  
<http://www.edict.com.hk/ConcordanceIndex/Business/default.htm>

WordSmith Tools website:  
<http://www.liv.ac.uk/~ms2928/homepage.html>

Ball, F (2001): Using corpora in language testing, *Research Notes* 6, 6-8

Hewings, M & Nickerson, C (eds) (1999): *Business English: Research into Practice*, Harlow: Longman

Hindmarsh, R (1980): *Cambridge English Lexicon*, Cambridge: Cambridge University Press

Van Ek, J A & Trim, J L M (1991): *Waystage 1990*, Strasbourg: Council of Europe Publishing

Van Ek, J A & Trim, J L M (1991): *Threshold 1990*, Strasbourg: Council of Europe Publishing

# The effect of training and standardisation on rater judgement and inter-rater reliability

STUART SHAW, RESEARCH AND VALIDATION GROUP

## Introduction

Examiner training is generally accepted as being essential to reliability and validity in the testing of second language performance (Alderson 1995:105); furthermore, training may play an important role in the professionalisation of language testing which has been called for by Bachman (2000:18). It is precisely for this reason that UCLES EFL already invests considerable resources and expertise in the initial training and ongoing standardisation of writing examiners for all the Cambridge EFL tests.

According to Weigle (1994:199), however, there has been little empirical research that might inform the development of effective training programmes. This article reports on an experimental study carried out recently at UCLES designed to investigate the effect of the training and standardisation process on rater judgement and inter-rater reliability. This study is part of a much larger and ongoing research programme to deepen our understanding of rater behaviour and to refine our existing approaches to rater training and standardisation.

## Purpose of the study

This study – undertaken in the context of the Writing Revision Project for CPE (Certificate of Proficiency in English) – focusses on the training and standardisation process as the variable most critical to improving the assessment of writing; it aims to empirically determine inter-rater reliability as well as deduce ways of improving inter-rater agreement. Specifically, the research questions for the study include:

- i. Does an iterative standardisation procedure improve the inter-rater reliability of multiple raters rating the same set of scripts?
- ii. What change is there during successive iterations of the standardisation in the scores given by raters?
- iii. How many iterations produce the best result?

## Revised CPE Paper 2 – Writing

The revised CPE Writing paper is based on realistic tasks with real world applications. As such, the nature of the writing tasks is defined as precisely as possible, with each task having the role of the writer, the role of the reader and the purpose for writing clearly defined. The range of tasks is defined to encourage candidates to develop a wide range of relevant writing skills within appropriate formats. The revised CPE Writing consists of two parts and candidates are required to carry out two tasks:

- Part 1 (Question 1) – a compulsory task;
- Part 2 (Questions 2-5) – candidates choose one task from a choice of four. Question 5 has one task on each of three set texts.

Candidates are expected to write between 300-350 words for each task in two hours.

In Part 1, candidates are asked to write within the following formats: an article, a proposal, an essay, and a letter. All the questions in this part have a discursive focus – presenting and developing arguments, expressing and supporting opinions and evaluating ideas – and are contextualised in order to provide guidance to the context through instructions and one short text which may be supported by visual prompts.

In Part 2, candidates are offered a choice of tasks within any of the following formats: an article, a letter, a proposal (not for set texts), a report, a review and an essay (set texts only). Each of the optional questions is a contextualised writing task specified in no more than 70 words. Candidates are expected to demonstrate the ability to write using a range of functions including narrating, analysing, hypothesising, describing, giving reasons, persuading and judging priorities.

The compulsory question in Part 1 provides a reliable means of assessing all candidates on the basis of one, standardised task and

gives all candidates an equal opportunity to produce a representative sample of their written work. The discursive focus is particularly relevant to students in education and for the academic application of CPE. The range of task types and topics in Part 2 allows candidates to select an optional question which is most relevant and interesting to them.

Responses are assessed using both a general mark scheme, which is used for all the questions, and a task specific mark scheme for each question. The criteria used to assess the candidates' answers in the general mark scheme include:

- Range of structure, vocabulary and expression appropriate to the register;
- An ability to organise content;
- An ability to write effectively and accurately, incorporating all aspects of the task.

Candidates need to meet the requirements set out in the task specific mark scheme before they can achieve the minimum acceptable performance for a CPE candidate.

## Research design

(It should be noted that the following description of the research design for the trial rating process does not reflect exactly what happens in the current live marking situation.)

In outline, the procedure was to:

- train a group of experienced Assistant Examiners (AEs) at a face-to-face meeting – using the new mark scheme;
- do multiple marking of a set of scripts at the same meeting for standardisation purposes;
- and then, off-site, do a series of iterations in which further sets of scripts were marked.

The marking process is shown diagrammatically in Figure 1.

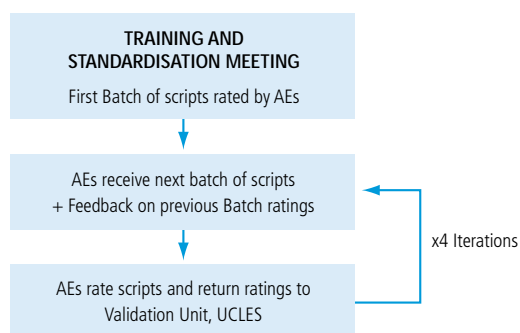


Figure 1 : Trial rating process

The scripts were taken from the May 2000 trialling of revised CPE Paper 2 tasks with candidate details removed to avoid any possible examiner bias. Raters used the new general mark scheme and the task specific mark scheme for each question to award an appropriate score. Both before training and standardisation and on 4 successive occasions over two months after the meeting, the AEs rated batches of scripts. This particular meeting permitted a

hierarchical style of co-ordination of marking as opposed to a consensual style. Each batch of marking (or iteration) following the meeting was preceded by a standardisation exercise pack consisting of the next batch of scripts to be marked, instructions, mark record sheets, task specific mark schemes and explicit feedback notes on each script in the batch previously marked explaining why the given mark was correct.

The given mark had been agreed by two CPE Paper 2 Principal Examiners (PEs) who had also provided feedback notes on each script. The individual scores from each PE for every script were collected and compared and an agreed 'standard' mark arrived at through consultation. The method employed for arriving at PE agreement was simple: if a response elicited a 2.1 and a 2.3 from each of the two PEs, an average was computed without further discussion. Scores which accounted for a greater difference were considered more carefully.

The apportionment of scripts and the timings of the project in relation to the raters' training and standardisation meeting are tabulated in Table 1.

The scores given by raters were recorded using the revised mark scheme. Each piece of writing is assigned to a band between 0 and 5 and can be awarded one of three performance levels within that band. For example, in Band 4, 4.1 represents weaker performance within Band 4; 4.2 represents typical performance within Band 4; 4.3 represents strong performance within Band 4. 'Acceptable' performance at CPE level is represented by a band of 3.

The scores given by AEs were compared with the standard ratings, as agreed by the PEs, for the same scripts, by subtracting the latter from the former. Thus, if an AE gave a particular script a score of 2.2, and the standard band score for that script was 2, the difference would be noted as zero; if an AE gave a score of 3.1 and the standard was 4, the difference was noted as minus 1; if an AE gave a score of 5.2 and the standard was 3, the difference was noted as + 2, and so on. The frequency with which the difference was zero, or -1, or +3, etc., was counted for each rater for each iteration, for both compulsory question 1 and optional question 2.

## Results

Table 2a and 2b summarise the overall percentage scores awarded by examiners for both the compulsory and optional questions over the five iterations.

As a whole, the gain in standardisation of rating over the first four iterations is not striking: the number of ratings 'On Standard' rose between IT1 and IT3 by nearly 10% to 58.9% for the 'compulsory' question; and by just over 9% to 46.8% between IT1 and IT4 for the 'optional' question. The percentage 'On or within one band of Standard' also rose from 92.3% to 96.1% for the compulsory task and 86.3% to 93.5% for the optional task.

## Inter-rater reliability

Multiple marking by a number of AEs using the same scripts provides a large number of inter-rater correlations, which are of

Table 1 : Timetable for data collection (timing of project iterations in relation to stages of the process) and script number apportionment

Standardising Exercise		Timing	Number of Scripts In Batch	Script Batch Number
Initial Training/Standardisation	(IT1)	May 16th, 2001	10	1
2nd Iteration	(IT2)	1 Week after Initial Training	25	2
3rd Iteration	(IT3)	2 Weeks after Initial Training	25	3
4th Iteration	(IT4)	4 Weeks after Initial Training	25	4
5th Iteration	(IT5)	6 Weeks after Initial Training	10	5
<b>TOTAL</b>		<b>8 Weeks</b>	<b>95</b>	<b>5</b>

Table 2a : Percentage of examiner scores in relation to Standard for the compulsory question

	-3	-2	-1	0	+1	+2	+3
<b>IT1</b>	0	5.13	29.91	48.72	13.68	2.56	0
<b>IT2</b>	0	3.53	23.72	46.47	20.51	5.49	0.85
<b>IT3</b>	0	1.28	21.79	58.97	15.38	1.28	1.28
<b>IT4</b>	0	–	–	–	–	–	–
<b>IT5</b>	0	3.85	39.74	39.74	11.54	5.13	0

– No compulsory question in Batch 4

Table 2b : Percentage of examiner scores in relation to Standard for the optional question

	-3	-2	-1	0	+1	+2	+3
<b>IT1</b>	1.07	8.55	23.93	37.61	24.79	3.42	0
<b>IT2</b>	0	2.99	23.50	40.60	23.08	8.12	1.28
<b>IT3</b>	0	0.85	32.05	46.58	16.24	4.27	0
<b>IT4</b>	0	0.92	17.23	46.77	29.54	5.54	0
<b>IT5</b>	3.85	9.62	32.69	36.54	13.46	3.85	0

interest in our research. To compute the inter-rater reliability for multiple raters for each of the five iterations a Pearson correlation matrix was generated and then an average of all the correlation coefficients was derived. Any distortion inherent in using the Pearson for ordinal data was corrected for by applying a Fisher Z transformation to each correlation. Tables 3 and 4 show inter-rater reliabilities for AEs and for PEs respectively.

Table 3 : Assistant Examiner Inter-rater Reliability on five occasions

Training/Standardisation Meeting	Iteration	Number of Scripts	Inter-rater Reliability
Pre-meeting	IT1	10	0.77
Post-meeting	IT2	25	0.77
Post-meeting	IT3	25	0.75
Post-meeting	IT4	25	0.75
Post-meeting	IT5	10	0.75

Table 4 : Principal Examiner Inter-rater Reliability on five occasions

Training/Standardisation Meeting	Iteration	Number of Scripts	Inter-rater Reliability
Pre-meeting	IT1	10	–
Post-meeting	IT2	25	0.906
Post-meeting	IT3	25	0.756
Post-meeting	IT4	25	0.604
Post-meeting	IT5	10	0.846

– No data collected

The AE inter-rater reliability is very constant varying by only 0.02. The PE reliabilities, however, are more erratic. With the exception of IT4, the estimates of reliability are high. When the five data sets, corresponding to the five iterations, are combined to form one set, the total computed inter-rater reliability for the AEs is 0.77.

## Discussion

This study focused on the training and standardisation process as the variable most critical to improving the assessment of writing and aimed to find ways of improving inter-rater agreement; it tested the hypothesis that a steady improvement in inter-rater correlation would take place with each successive iteration of the standardisation exercise. However, results reveal that whilst the inter-rater reliabilities are encouragingly high, they do not improve with time and standardisation but remain roughly constant.

Interestingly, the data shows evidence of examiners modifying their behaviour with successive standardisation exercises. The columns in Figure 2 represent percentage of examiner scores in relation to 'standard' scores for examiners marking 'lower than standard', 'on-standard' and 'higher than standard' for the optional question. The scores by the raters in IT1, i.e. before training and standardisation, do not differ grossly from the standard. Initial results may well reflect examiner experience despite the fact that half the AEs were unfamiliar with the revised mark scheme. It is possible that the mark scheme, comprising a set of detailed and explicit descriptors, engenders a standardising effect even in the absence of a formalised training programme. The group had a tendency to harshness with roughly equal severity on the compulsory and optional questions although the examiners were nearly twice as generous on the optional question. The mark scheme applied to the compulsory question is both more rigid and more clearly defined than its optional question counterpart. Additionally, the range of language required by the compulsory task is less wide and its focus is discursive whereas the optional task permits more scope for invention and a variety of interpretation. Consequently, examiners are allowed greater freedom in their assessment of the optional response which may account for increased leniency.

The evidence from the scores for IT2 suggests that standardisation prompted some adjustment in the severity of examiner rating. There was a trend to increased leniency. The group rated significantly less severely in IT2 which may be a consequence of the greater attention given to the revised mark scheme. For both tasks, the group were more generous in their awards. As far as changes in relative severity/leniency are concerned, the results of this study are broadly in line with Weigle's finding that experienced raters are more generous than perhaps inexperienced ones (1998:263). 'On Standard' scores show a marginal decrease for the compulsory task and a slight increase for the optional question.

Significant improvement is manifest for IT3 for both 'On Standard' and 'Within +/- One Band of Standard' for both the compulsory and optional questions. For the compulsory task, examiners are less harsh and less lenient than for IT2. However, more interestingly, examiners assessing the optional question reversed the trend of IT2 with more generous marking. A pattern is beginning to be established which reflects alternating trends between low and high marking over the various standardisations creating a 'see-saw' effect.

IT4, including a batch of scripts which constituted optional questions only, reinforced the emerging 'see-saw' pattern. The percentage of ratings 'On Standard' remained roughly constant and the percentage 'Within +/- One Band of Standard' is virtually unaltered. However, a significant shift from harshness towards leniency is manifest, reflecting the earlier trend at IT2. Over the first four iterations, the percentage of aberrant ratings i.e. more than one band from standard, fell for both compulsory and optional questions.

The results for IT5, however, are erratic. Batch 5 consisted of only 10 scripts and were a collection of different tasks: Revision

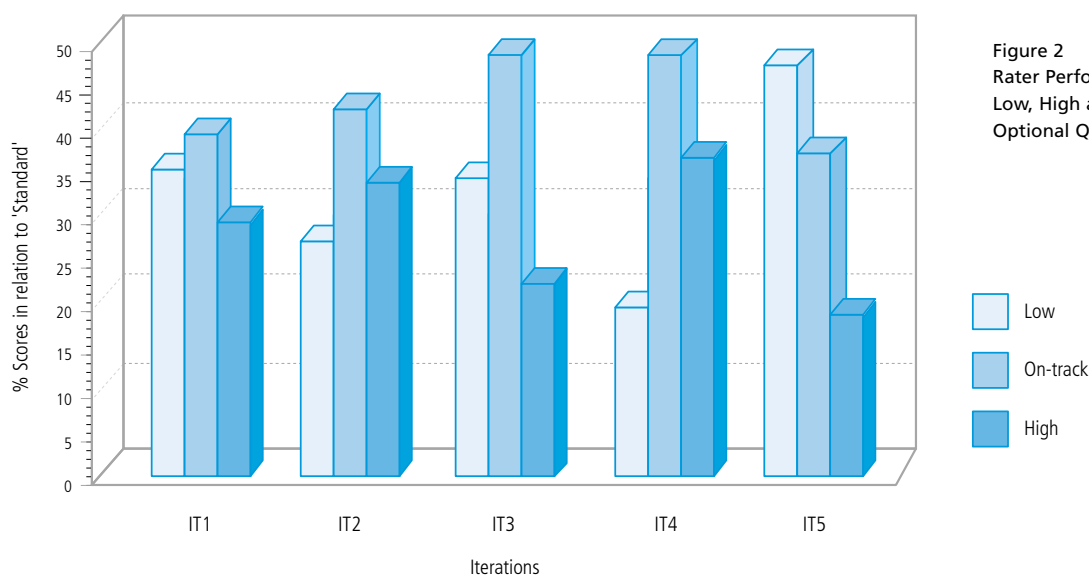


Figure 2  
Rater Performance over 5 iterations:  
Low, High and 'On-Track' Marking  
Optional Question



Task B, Task M and Set Text 2. Moreover, the marking of Batch 5 scripts coincided with 'live' marking of the June administration. It may be that examiners at this point in the trial were experiencing 'participation fatigue' and 'divided loyalty'.

Despite the fact that more examiner ratings are increasingly 'on-track', the extent of examiner over-compensation appears to be increasing as the trial continues. It would appear that some examiners were becoming increasingly concerned by their lack of consistency with 'standard' ratings.

According to interviews conducted with AEs after the trial, examiner confidence throughout was affected in varying degrees. Many examiners were worried by the frequency with which they appeared to be 'off-track' when their ratings were compared with 'standard' ratings, especially when their ratings were greater than one band score from 'standard'. Discrepancies were thought to be related to training issues and rater variation attributed to limited training opportunities with the revised mark scheme. Whenever it was perceived that AEs were 'off-track' some corrective action was considered. The nature of this action was symptomatic of the extent of any variation and examiner personality. Examiners were provoked into taking a range of adjustments to their individual assessment approach. For certain examiners, however, no adjustments were made.

## Future research

A principal research consideration regarding the assessment of subjective tests is to ascertain ways to improve the reliability of the marking of writing. Studies like this one demonstrate the value of investigating approaches to rater training and standardisation in relation to the use and appropriate application of mark schemes.

Further research is currently being undertaken which will address issues relating to the nature of feedback to examiners and the effect of feedback on examiners. The quality of feedback to the examiner is likely to be an important factor in the success, or otherwise, of training and standardisation. Wigglesworth (1993) found evidence that examiner bias was reduced following feedback and that examiner-rater consistency improved. These issues include the role of Team Leader feedback in examiner training and standardisation, its extent and the form. With regard to mark scheme criteria, the extent to which examiners consistently pay attention to the relevant criteria after training and standardisation and during the marking process also needs to be addressed.

Some researchers suggest that the views of examiners are not taken into consideration during the standardisation process (Pinot de Moira and Mac, 1999). Questions arising from the role of examiner discussion in standardising/co-ordinating marking include:

- Does discussion between examiners produce more consistent marking in terms of accuracy and reliability?
- Are the views of examiners taken into account and to what extent?

- Does being included in what matters an essential requirement for being engaged in a 'community practice'?

Wolf (1995) implies that the exchange of viewpoints is important for facilitating a 'community of practice':

- Is a consensual style of co-ordination more beneficial than a hierarchical style in the co-ordination of marking?
- Is marker reliability improved the more the markers concerned form part of a group in constant contact and discussion with each other?
- Do aberrant examiners negatively influence the judgements of other examiners in discussions?

A future study might compare the more traditional, hierarchical style of marker co-ordination with a consensus style of co-ordination.

Finally, this study is also of relevance in relation to possible technological developments in the large-scale assessment of writing and their implications for marking reliability. Trials have recently been conducted into the feasibility of Electronic Script Management (ESM) for EFL examinations, including a trial investigating marking Paper 2 of the Certificate of Advanced English – CAE. Future issues of *Research Notes* will report on these and other studies in more detail.

## References and further reading

- Alderson, J C, Clapham, C and Wall, D (1995): *Language test construction and evaluation*, Cambridge: CUP
- Bachman, L F (2000): Modern language testing at the turn of the century : assuring that what we count counts, *Language Testing* 17/1, 1-42
- Bachman, L F, Davidson, F, Ryan, K and Choi, I (1989): *The Cambridge-TOEFL Comparability Study : Final Report*, Cambridge : University of Cambridge Local Examinations Syndicate
- Lumley, T and McNamara, T F (1995): Rater characteristics and rater bias : implications for training, *Language Testing*, 12/1, 54-71
- Pinot de Moira, A and Mac, Q (1999): Survey of examiner views. AEB Internal Report RAC/799
- Weigle, S W (1994): Effects of training on raters of ESL compositions, *Language Testing*, 11/2, 197-223
- Weigle, S C (1998): Using FACETS to model rater training effects, *Language Testing* 15/2, 263-287
- Wigglesworth, G (1993): Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction, *LanguageTesting*, 10/3, 305-335
- Wolf, A (1995): *Competence-based assessment*. Open University Press: Bristol

## Review of recent validation studies

### Investigating gender differences in young learner performance

A recent study into performance on the Cambridge Young Learner Tests was completed just too late to be included in our last *Research Notes* with its special focus on YLE. An analysis of score data for almost 60,000 YLE candidates who have taken the tests since they were introduced in 1997 produced the following overall results:

- at *Starters* levels girls appear to achieve significantly higher scores than boys on all three components – Reading/Writing, Listening, and Speaking
- at *Movers* and *Flyers* levels, girls appear to score significantly higher scores than boys on the Reading/Writing and Listening components
- at *Movers* and *Flyers* levels boys tend to achieve higher scores than girls on the Speaking component (though the difference is only significant for *Flyers*)

### Investigating test conditions for listening and speaking

UCLES is committed to ensuring that all candidates are treated fairly and have an equal opportunity when taking any of our tests. In order to confirm this, we routinely monitor candidate and test performance at each administration to check for any unexpected differences in performance; in addition, we sometimes carry out special studies to investigate specific questions. Two recent internal studies set out to answer the following questions:

1. Does using CDs (as opposed to cassettes) to administer the Listening paper have any effect on performance?
2. Does taking the Speaking test on the same day as the other components (as opposed to a different day) have any effect on performance?

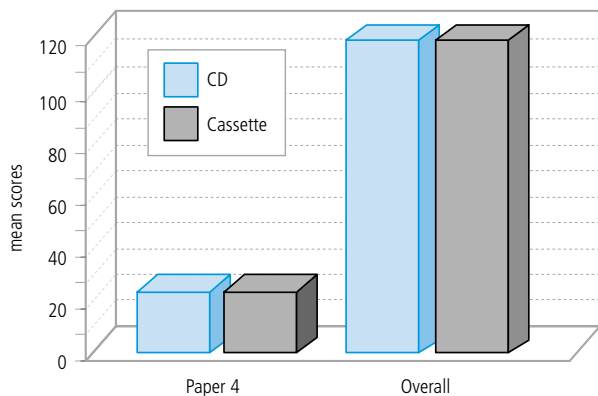
#### Study 1

This study investigated candidate performance in the Listening Paper for June 2001 and December 2001 administrations of FCE Syllabus 0101, taken by over 97,000 candidates in Greece.

The option of using CDs instead of cassettes to deliver the listening test was first offered by UCLES EFL in 2000; this option had been requested by many Cambridge centres and it was only introduced after a successful trialling exercise in 1999 in various parts of the world. In principle, using a CD rather than a cassette results in a better quality of recording so some people have been concerned that candidates might be advantaged/disadvantaged according to whether they are listening to a CD or cassette.

In June 2001 less than 10% of the total candidature in Greece used the CD for the Listening paper (Paper 4), while in December this figure rose to just over 60%. A comparison of Paper 4 and overall mean scores across both test administrations showed no significant difference in results between candidates hearing the Listening test via CD and those hearing it via cassette (see Figure 1 below).

Figure 1: CD vs cassette: mean scores on paper 4 and overall



In other words, the study found no evidence to suggest that performance on the Listening Paper (Paper 4) is affected by mode of delivery; so teachers, parents and students can be reassured that candidates are being treated fairly whichever listening option they encounter.

#### Study 2

A second study set out to investigate whether taking the Speaking test on the same day as other components (i.e. Reading, Writing and Use of English) might mean that candidates underperform, possibly as a result of being tired after having taken the written papers earlier in the day. If fatigue were a factor, then one might expect to see lower mean scores for candidates taking their Speaking test on the same day as the written components when compared with the scores of those who took the test on a different day. (Candidates normally take their Speaking test on one day within a window period of 4-5 weeks.)

The scores of FCE, CAE and CPE candidates who took the written components (Papers 1, 2 and 3) on the same day as their Speaking test were compared with the scores of those who took the speaking test on a different day; and a comparison was also made with the overall mean Speaking test score for that test session. The analysis included data from test administrations which took place in March, June and December between 1999 and 2001. A review of scores from 14 test administrations for FCE, CAE and

CPE suggested that candidates who take their Speaking test on the same day as the written components perform just as well as those who take it on another day within the window, whether before or after. The analysis found no evidence of lower scores due to fatigue. It may be that any fatigue that does result from having taken several written components before the Speaking test is

counterbalanced, or even outweighed, by the adrenalin flow on the day, or by the benefit of doing the written papers beforehand! Once again, test users can be reassured that, whatever day they take the Speaking test, candidates have an equal opportunity to demonstrate their spoken language ability.

## Other news

### QCA Accreditation

The UK government's exams regulator QCA has accredited the Cambridge EFL examinations as part of the UK's National Qualifications Framework (NQF). Accreditation currently covers the following exams:

NQF level	Examinations		
3	CPE		
2	CAE	BEC Higher	CELS Higher
1	FCE	BEC Vantage	CELS Vantage
Entry 3	PET	BEC Preliminary	CELS Preliminary
Entry 2	KET		

An IELTS score of between 6 and 7 has also been accredited at level 2.

The teacher certification examinations – CELTA and DELTA – have been submitted for accreditation and are expected to be accredited at levels 4 and 5.

In the next few months, we will be considering how best to make use of this accreditation in the UK and other countries. It would be useful to hear how this accreditation could have beneficial impact on the value of certificates in your country. Please email your comments to Stephen McKenna, [mckenna.s@ucles.org.uk](mailto:mckenna.s@ucles.org.uk)

More details are available at [www.cambridge-efl.org/QCA](http://www.cambridge-efl.org/QCA)

### Access to internal reports

We frequently receive requests from external researchers asking for access to some of the reports we produce in the course of our research/validation work and which are often referred to at the end of articles in *Research Notes*. Unfortunately, it is usually not possible for us to provide access to this material.

Although we undertake a great deal of research related to our tests, most of this tends to be written up for internal, operational purposes only. In certain cases, we try to produce final reports of some of our studies for publication in the wider domain but this process inevitably requires a considerable amount of time and other resources and only a limited number of studies can be published in this way.

*Research Notes* is one way in which we are able to report on some of our work for the benefit of the research community and the wider public, though it is sometimes difficult for us to do this at the level of detail which external researchers and students would like to see. Later this year, the forthcoming volumes in the *Studies in Language Testing* series will be able to provide far more detailed information on the wide range of studies which have contributed to the recent revisions of CPE and BEC, and to the development of CELS.

### New Candidate Information Sheets

A new Candidate Information Sheet has recently been introduced for all examinations. The Candidate Information Sheet is filled in by candidates at each examination session; the form gathers valuable information about candidates' linguistic and demographic backgrounds and their reasons for taking the test. Changes have been made to Question 3, which combines two previous questions about study and work, and Question 7, where new and revised examination names have been added. One new question has been added which asks candidates who have been sponsored to tell us the name of the sponsoring company.

The Candidate Information Sheets help us maintain the relevance of the examination tasks and avoid bias. All information collected is covered by the UK Data Protection Act.

## Conference reports

In February 2002 BALEAP (British Association of Lecturers in English for Academic Purposes) held a Professional Issues Meeting on the subject of 'Accuracy in EAP' at the School of Oriental and African Studies, London University. A member of the Research and Validation staff contributed a 90-minute presentation/workshop session to this event entitled 'Writing performance, IELTS and issues in assessing accuracy'. The session included a brief review of the format and content of the IELTS Academic Writing Module; the criteria used for assessing candidates' writing were considered, in particular how the notion of accuracy is conceptualised and operationalised for assessment purposes in IELTS and in other English language proficiency tests. Findings from recent studies of candidate and examiner performance were also presented, together with examples of writing scripts at different levels for discussion purposes.

The 9th International House Symposium was held in Torres Vedras, Portugal in March 2002. This event was aimed at teachers and trainers and its theme was current theory and practice in English language teaching. Several UCLES personnel attended this symposium and used the opportunity to develop a better understanding of English teaching in the Portuguese context. The contribution to the symposium from Research and Validation was a presentation on 'Current Perspectives of Corpus-Informed Language Testing'. This talk outlined UCLES' current use of native speaker and learner corpora for three purposes: as an archive of examination scripts, for research into speaking and writing, and for operational activities to support all of UCLES examinations. Several current research projects were also described, including the corpus of Young Learner Speaking Tests (see *Research Notes 7*, page 8) and the development of item-writer wordlists (see article on

page 10 in this issue). The presentation also considered what teachers can do with corpora which prompted interest in the predominantly teacher audience. A view of the future use of corpora in teaching and testing was also provided. This presentation intended to demonstrate that links can and should be maintained between examination boards and teachers. The Symposium is a biannual event and is of continuing relevance to UCLES as an examination provider. (See International House website <http://www.ihworld.com/>)

Another event held during March 2002 was a Symposium on Assessing Intercultural Competence at the School of Education, University of Durham (UK). Around 30 people attended the symposium representing 15 countries in Europe and North America; the purpose of the event was to develop an informal 'conversation' over a 3-day period on the nature of intercultural competence and its potential for formal/informal assessment. Sessions included: a survey of existing approaches and techniques; reports on specific projects in different countries; discussions on levels and grading and on the ethical issues associated with assessing attitudes. Contributors were able to share their insights and experience in what is a relatively new and undeveloped field for language specialists; plans are already in hand to build on this initial contact and to identify an agenda for further development and possible research.

The conference season became particularly busy towards the end of March and in early April with IATEFL in York (UK) and with AAAL and TESOL in Salt Lake City (USA) and there will be a report on contributions by Research and Validation staff to these key conferences for English language teaching/testing in Issue 9 of *Research Notes*.

## Studies in Language Testing – Volume 15

Volume 15 in the *Studies in Language Testing* series documents in some detail the most recent revision of the Certificate in Proficiency in English (CPE) which took place from 1991 to 2002. CPE is the oldest of the Cambridge suite of English as Foreign Language (EFL) examinations and was originally introduced in 1913. Since that time it has been regularly revised and updated to bring it into line with current thinking in language teaching, applied linguistics and language testing theory and practice.

For many years, much of the work that took place behind the scenes at UCLES remained fairly obscure to users of Cambridge EFL examinations around the world. However, in recent years there has been a serious attempt to inform users more effectively about

what UCLES does and how it does it. Increased information has come in a variety of ways including: regular meetings with Local Secretaries (the official in-country providers of Cambridge EFL examinations) all over the world; a comprehensive programme of teacher seminars focusing on test content and candidate performance; regular newsletters such as *Cambridge First* and *Research Notes*; involvement in international language testing groups and associations such as the Association of Language Testers in Europe (ALTE); and frequent presentations at local and international conferences.

The publication of Volume 15 is a further illustration of UCLES' desire to provide users of its EFL examinations with an in-depth

understanding of what it does and how it operates by making the thinking, processes and procedures that underpinned the current revision of CPE as explicit as possible. The volume also seeks to provide an honest account of the revision process, the questions and problems faced by the revision teams, the solutions they came up with and the challenges that face UCLES EFL in the future. The volume is intended to be of interest and relevance to a wide variety of readers. For those interested in a historical perspective, Chapter 1 traces the history of CPE from its first version in 1913 through to the present day and beyond. For those interested in how UCLES works, Chapter 2 documents in some detail the test development and production process used in relation to the CPE and its revision as well as in a more general sense. Chapters 3-7 provide detailed information for those interested in why the papers look the way they do, what went into designing, piloting and confirming their final characteristics as well as insights into the writing of various materials. Finally, Chapter 8 looks to the future. The work of an examination board is never done. When one revision finishes another begins and so it is with CPE. The volume is a true team effort, as is so much of the work done by UCLES

EFL. The chapters are written by seven different authors and commented on by a number of other individuals. The work reported involves many teachers, candidates, consultants, examiners, subject officers and others.

Unfortunately, work in public examinations has tended to be ephemeral and few accurate or comprehensive records are easily accessible. It is hoped that this volume will begin to reverse that pattern and it is to be followed soon by three further volumes each documenting a revision process and providing a historical context for the examinations in question. The Certificates in English Language Skills were launched in May 2002 so in Volume 16 Roger Hawkey, working with a team of UCLES EFL subject officers, traces the history of several examinations that were withdrawn with the introduction of CELS but which have played a part in its evolution. In a later volume Barry O'Sullivan, also working with UCLES staff, documents the revision of the Business English Certificates and Alan Davies is currently working on tracing the evolution of tests of academic English with particular reference to the development of IELTS.

## IELTS joint-funded research 2002 (Round 8): call for proposals

All IELTS-related research activities are co-ordinated as part of a coherent framework for research and validation. Activities are divided into areas which are the direct responsibility of UCLES, and work which is funded and supported by IELTS Australia and the British Council.

As part of their ongoing commitment to IELTS-related validation and research, IELTS Australia and the British Council are once again making available funding for research projects in 2002. For several years now the two partners have issued a joint call for research proposals that reflect current concerns and issues relating to the IELTS test in the international context (see article below). Such research makes an important contribution to the monitoring and test development process for IELTS; it also helps IELTS stakeholders (e.g. English language professionals and teachers) to develop a greater understanding of the test.

All IELTS research is managed by a Research Committee which agrees research priorities and oversees the tendering process. In determining the quality of the proposals and the research carried out, the Committee may call on a panel of external reviewers. The Committee also oversees the publication and/or presentation of research findings.

### What areas of interest have been identified?

At its last meeting, the IELTS Research Committee identified the following as among the areas of interest for research purposes:

- work relating to the revised IELTS Speaking Test (e.g. investigation of examiner/candidate discourse, study of examiner/candidate attitudes to the revised format);
- work relating to the range of tests now used for university/college entry in Australia/New Zealand/UK/Canada, including methods/criteria used by university admissions staff and faculty heads when deciding acceptable English language thresholds for their courses;
- work relating to IELTS and test impact (e.g. a study of the development and use of IELTS preparation courses, IELTS course materials, the attitudes of IELTS stakeholders);
- work relating to band score gain and intensive English language training, including the recommended language threshold below which students should not attempt an IELTS test;
- work on other issues of current interest in relation to IELTS.

### Is access to IELTS test materials or score data possible?

Access to IELTS test materials or score data is not normally possible for a variety of reasons, e.g. test security, data confidentiality. However, sometimes a limited amount of retired material (e.g. writing test prompts) may be made available for research purposes. In addition, UCLES has been engaging over recent years in the development of instruments and procedures designed to investigate the impact of IELTS; it is possible that these may be made available for use by researchers following consultation with UCLES (more details are given in the IELTS Annual Review 2000/2001).

### Who may submit proposals?

As part of the IELTS policy of stimulating test-related research among its stakeholders, it is hoped that many of the research proposals in 2002 will be submitted by researchers and organisations who have a connection with IELTS, e.g. consultants, Senior Examiners, IELTS Administration Centres and centres which have assisted in trialling IELTS. There is, however, no objection to proposals being submitted by other groups/centres/individuals.

### What is the level and duration of funding available?

The maximum amount of funding which will be made available for any one proposal is £13,000/AUS\$30,000. The research study will need to be completed and a full report submitted by the end of December 2003.

### What is the procedure for submitting proposals?

Proposals for funding should take the form of a typed/word-processed document of no more than 10 pages and should be accompanied by a completed application form and its attachments (available from the addresses given below).

### Who will evaluate the proposals?

All research proposals will be evaluated by the IELTS Research Committee comprising representatives of the three IELTS partners as well as other academic experts in the field of applied linguistics and language testing.

### What criteria will be used to evaluate proposals?

The following factors will be taken into consideration when evaluating proposals:

- Relevance and benefit of outcomes to IELTS
- Clarity and coherence of proposal's rationale, objectives and methodology
- Feasibility of outcomes, timelines and budget (including ability to keep to deadlines)
- Qualifications and experience of proposed project staff
- Potential of the project to be reported in a form which would be both useful to IELTS and of interest to an international audience

### What is the time scale for the submission and evaluation of proposals?

The following time scale will apply:

<b>31 July 2002</b>	Deadline for submission of proposals
<b>August/September 2002</b>	Preliminary review of proposals by IELTS partners
<b>October/November 2002</b>	Meeting of IELTS Research Committee to evaluate and select successful proposals

### Application forms and submission guidelines are available from:

#### Ms Sasha Hampson

Program Manager  
Testing Services  
IELTS Australia  
IDP Education Australia  
GPO Box 2006  
Canberra  
ACT 2601  
Australia

Tel: 61 6 285 8222

Fax: 61 6 285 3233

E-mail: [sasha.hampson@idp.com](mailto:sasha.hampson@idp.com)

#### Ms Helen Bird

UK and Ireland IELTS Manager  
IELTS Research  
British Council  
14 Spring Gardens  
London  
SW1A 2BN  
United Kingdom

Tel: 44 20 7389 4726

Fax: 44 20 7389 4140

E-mail: [helen.bird@britishcouncil.org](mailto:helen.bird@britishcouncil.org)

[www.ielts.org](http://www.ielts.org)

## IELTS joint-funded research programme – 1995-2001

For several years now, IELTS Australia and the British Council have issued a joint call for research proposals reflecting current concerns and issues relating to the IELTS test in the international context. The idea of allocating funding for external research into IELTS dates back to 1995 when IELTS Australia first decided that additional qualitative information would be valuable on a range of issues, particularly those affecting the continuing reliability and standing of the test. The activity of selecting worthwhile research proposals from the large number of submissions received began in

the same year and continued for three years. In 1998 the British Council decided to allocate similar funding for external research into IELTS and since that time the two partners have issued joint calls for research proposals. As the third partner in IELTS, UCLES EFL often supports the IELTS/BC-funded projects by providing relevant information, materials or data. Since 1995 more than 40 IELTS-related research projects and nearly 60 different researchers have received funding under this programme (a full list is given below).

## IELTS related research projects funded by IELTS Australia/British Council

Round/Year	Topic	Researchers
<b>One/1995</b>	Survey of receiving institutions' use and attitude towards IELTS	Clare McDowell & Brent Merrylees
	Comparison of writing assessment procedures	Greg Deakin
	An investigation into approaches to IELTS preparation with a particular focus on the Academic Writing component of IELTS	James D H Brown
	A comparative study of IELTS and Access test results	Magdalena Mok
	The effect of interviewer behaviour on candidate performance in the IELTS oral interview	Alan Davies & Annie Brown
	The misinterpretation of questions in the reading and listening components of the IELTS test	Stephen Heap & Gayle Coleman
	An investigation of the predictive validity of IELTS amongst a sample of international students at University of Tasmania	Fiona Cotton & Frank Conrow
<b>Two/1996</b>	A comparison of IELTS and TOEFL as predictors of academic success	Brian Lynch, Kathryn Hill & Neomy Storch
	Construct validity in the IELTS Academic Writing Module: a comparative study of Task 2 topics and university writing assignments	Tim Moore & Janne Morton
	IELTS in context – issues in EAP for overseas students	Robynne Walsh & Greg Deakin
	Specifying the internal and the candidate group profiles of IELTS results in 1996 from Australian test centres	A. Lee, Christine Bundesen & Magdalena Mok
	An investigation of the effect of students' disciplines on their IELTS scores	Cynthia Celestine
<b>Three/1997</b>	An investigation of speaking test reliability with particular reference to candidate/examiner discourse produced and examiner attitude to test format	Clare McDowell & Brent Merrylees
	The relevance of IELTS in assessing the English language skills of overseas students in the private education and training sector	Greg Deakin & Sue Boyd
	The impact of gender in the IELTS oral interview	Kieran O'Loughlin
	A study of response validity of the IELTS writing module	Carol Gibson, Peter Mickan & Stephan Slater
	An investigation of raters' orientation in awarding scores in the IELTS oral interview	Annie Brown
	Predictive validity in the IELTS test; a study of the relationship between minimum IELTS scores and students' academic success	Mary Kerstjens & Caryn Nery
	Monitoring IELTS examiner training effectiveness	Clare McDowell
<b>Four/1998</b>	A monitoring program of examiner performance in IELTS Australia centres	Brent Merrylees
	An evaluation of selected IELTS preparation materials	Judy Coleman & Rae Everett
	An impact study of 2 IELTS user groups: immigration and secondary	Brent Merrylees
	A study of the response validity of the IELTS Writing test- Stage two	Peter Mickan
	The validity of the IELTS test in an Open and Distance Learning (ODL) context	Elizabeth Manning and Barbara Mayor
	Impact study proposal	Dianne Schmitt
Identifying barriers in performance-based language tests in Korea	Young-Shik Lee and Peter Nelson	

Round/Year	Topic	Researchers
Five/1999	An analysis of the linguistic features of output from IELTS Academic Writing Tasks 1 and 2	Barbara Mayor, Ann Hewings & Joan Swann
	Investigation of linguistic output of Academic Writing Task 2	Chris Kennedy & Tony Dudley-Evans
	The effect of standardisation training on rater judgements for the IELTS Writing Module	Mark Rignall & Clare Furneaux
	Task design in Academic Writing Task 1: the effect of quantity and manner on presentation of information on candidate writing	Kieran O'Loughlin & Gillian Wigglesworth
	An investigation of the scoring of handwritten versus computer based essays in the context of IELTS Writing Task 2	Annie Brown
	The impact of the IELTS test on preparation for academic study in New Zealand	John Reed & Belinda Hayes
Six/2000	Monitoring score gain on the IELTS Academic Writing module in EAP programmes of varying duration	C.J. Weir & Antony Green
	Assessing the value of bias analysis feedback to raters for the IELTS Writing Module	Barry O'Sullivan & Mark Rignall
	Investigation of linguistic output of General Training Writing Task 2	Chris Kennedy
	What's your score? An investigation into performance descriptors for rating written performance	Peter Mickan
	Investigating the relationship between intensive EAP training and band score gain on IELTS	Catherine Elder & Kieran O'Loughlin
	The attitudes of IELTS stakeholders: administrator, lecturer and student perceptions of IELTS in Australian and UK universities	R.M.O. Pritchard, Roisin Thanki, Sue Starfield & David Coleman
	A comparative study of Academic IELTS and General Training IELTS for the secondary school market	Cynthia Celestine
Seven/2001	The impact of IELTS on the preparation classroom: stakeholder attitudes and practices as a response to test task demands	C.J. Weir & Antony Green
	Issues in the assessment of pen and paper and computer-based IELTS writing tasks	Russell Whitehead
	A longitudinal study of the effects of feedback on raters of the IELTS Writing Module	Barry O'Sullivan & Mark Rignall
	Assessing the impact of IELTS preparation programs on candidate's performance on the General Training Reading and Writing Module	Chandra Rao, Kate McPherson, Rajni Chand & Veena Khan
		A cross sectional and longitudinal study of examiner behaviour in the revised IELTS speaking test

The list above illustrates the broad range of issues and themes which have been addressed through the IELTS Australia/BC-funded research programme. Findings from many of these studies have helped to inform revisions to the IELTS test (e.g. the revised IELTS Speaking Test) and have helped shape other developments relating to IELTS (e.g. impact projects, market strategies).

Following completion of a research study, a final project report is submitted to the funding partner and is then reviewed by members of the Research Committee, which includes

representatives of all 3 IELTS partners. Before proceeding to publication, a report is refereed by one or more independent academic experts and revisions may be required by the authors before a report can be published.

IELTS Australia published some of the completed research projects from Rounds 1-3 in three volumes of *IELTS Research Reports* in 1998, 1999 and 2000 (see IELTS website for details). A further selection of completed reports will also appear in a volume in the *Studies In Language Testing* series (2002/3).