2

Research Notes

UNIVERSITY OF CAMBRIDGE LOCAL EXAMINATIONS SYNDICATE

ENGLISH AS A FOREIGN LANGUAGE (EFL)

ALTE Association of Language Testers in Europe

EAQUALS ASSOCIATE MEMBER

UNIVERSITY of CAMBRIDGE
Local Examinations Syndicate

ResearchNotes

Introduction

Welcome to the second issue of Research Notes, the newsletter about current developments in the research, validation and test development work carried out by the EFL Division at UCLES. We have received very positive feedback on the first issue, which gave an overview of the research carried out at UCLES. This issue will look at some of the projects introduced in the last issue in more detail, as well as introducing some new topics.

Lynda Taylor looks at the complex patterns of stakeholders in language testing – such as test-takers, test-users, teachers and official bodies – and how UCLES includes them in the whole assessment process, from the development of new tests, to provision of special arrangements for candidates who are unable to take the standard format of the tests because of illness or disability. Nick Saville also considers the role of stakeholders when he discusses the impact of international language examinations on the language testing constituency, in the context of the International English Language Testing System (IELTS).

Issue 1 introduced the EFL Local Item Banking System, which UCLES EFL uses to construct examination papers. Here Simon Beeston describes in more detail its role in assuring the quality of Cambridge EFL examinations.

Developing new test items is a highly complex process, which draws on research in the field of applied linguistics, as well as extensive trialling and consultation. David Booth and Nick Saville discuss the development of the new gapped sentences task in the revised CPE Paper 3.

Neil Jones describes the development and validation of the ALTE 'Can-do' statements, and their relationship to the Council of Europe's Common European Framework. The 'Can do' statements describe the skills a typical learner should have at each level, and, along with the Council's Framework, can be used to compare the levels of language examinations in a range of European languages.

The use of the paired format of two examiners and two candidates for the Speaking Test has been adopted in many of the Cambridge EFL examinations. Lynda Taylor examines the benefits of testing candidates' speaking skills in this way. Nick Saville continues the Speaking Test theme and looks at the development and use of observation checklists in the validation of the Speaking Tests in the Cambridge Main Suite examinations.

In the next issue of Research Notes, Simon Beeston will continue his look at the Item Banking System. There will also be articles on partial credit analysis, the development of the IELTS rating scale, computer based testing and the China Project.

Research Notes is intended to reach a wide audience of people involved in Cambridge examinations around the world and also people who are interested in the theoretical and practical issues related to language assessment. We would be very interested to hear your views on the newsletter – whether you find it interesting and useful, how appropriate you find the level of presentation and if there are any topics you would like us to cover.

Research Notes is being distributed to all UCLES EFL centres and other key contacts. If you would like to receive additional copies or would like a personal subscription to the newsletter, please complete and return the form on page 20.

Contents

Published August 2000

Stakeholders in language testing

Lynda Taylor, Performance Testing Co-ordinator, UCLES

Several different models have been presented to describe the relationships between the 'stakeholders' involved in the business of testing. The most traditional model usually separates stakeholders into the producer and the consumer, the test-maker and the test-taker. More extreme views offer a socio-political view of testing in which power is exercised by one party over another. Closer analysis tends to reveal a far more complex community of participants and set of relationships than is represented by the models described above. Rea-Dickins (1997), for example, identified at least 5 stakeholder categories: learners, teachers, parents, government and official bodies, and the marketplace; but even this list can be developed into a much broader conceptualisation (Figure 1).

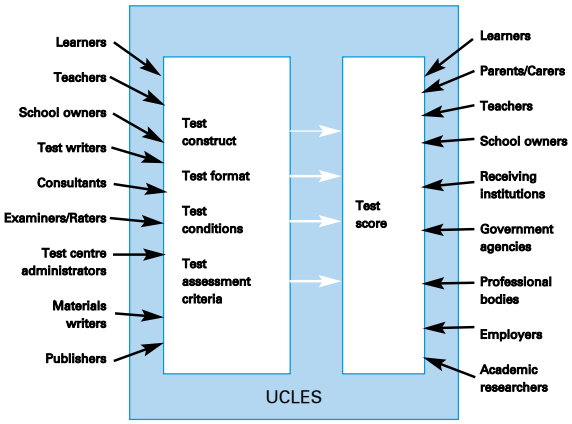

Figure 1

Some of the stakeholders listed above (e.g. examiners and materials writers) are likely to have more interest in the 'front end' of a test, i.e. the test assessment criteria or test format. Others may see their stake as being primarily concerned with the test score. Some stakeholders, such as learners and teachers, will naturally have an interest in all aspects of the test.

This article briefly describes five specific aspects of activity undertaken by UCLES in its role as a major test provider: the methodology for revising tests, the team leader system for oral examiners, the ALTE code of practice, support and information for stakeholders, and provision of special arrangements. Each of these areas of activity reflects the examination board's commitment to professional, ethical and legal accountability towards its fellow stakeholders.

1. Test revision methodology

Cambridge EFL examinations have been around for many years and have, of course, been revised on several occasions; FCE was last revised in 1996 and CPE is currently undergoing a revision process for introduction in 2002. Both projects can be used to illustrate how test revision is now part of an on-going validation process. As well as the routine analysis carried out on the current test formats, the process typically includes specially commissioned investigations and surveys. For example, in the 4 years leading to the 1996 revision of FCE, a user-survey administered questionnaires and structured group interviews to 25,000 students, 5,000 teachers and 1,200 oral examiners in the UK and around the world; 120 receiving institutions (universities and colleges) in the UK were also canvassed for their perspective.

As part of the current CPE Revision project, the revised draft materials have so far been trialled with nearly 3,000 candidates in 14 countries representative of the candidature world-wide. In addition, consultative seminars and invitational meetings have involved about 650 participants in eleven countries throughout Europe and South America. Feedback from all stages has been reviewed constantly and this has informed the progress of the revision at every stage.

The UCLES test revision process normally involves a number of interlinked stages, each following on from the next in a cyclical pattern. The process begins with research – this involves a number of specially commissioned investigations and market surveys as well as the routine analysis carried out on the current test format. After initial research a draft version of the test specifications is prepared. These are considered and reviewed by external consultants, each concentrating on a specific paper. The tasks are then redrafted and trialled on groups of candidates around the world.

Feedback from all stages is reviewed constantly and informs the progress of the revision at every stage, especially in reviewing specifications and materials. Consultative seminars and invitational meetings give teachers, Directors of Studies, teacher trainers and other agencies the chance to comment on the progress of the revision.

2. Team Leader System for Oral Examiners

Assessing spoken language performance, especially using a face-to-face test format, is complex because of the many variables involved and it is understandable that concerns relating to quality and fairness may be expressed by different stakeholders in the process.

There are currently over 10,000 approved UCLES EFL Oral Examiners (OEs) around the world involved in conducting one or more of the face-to-face Speaking Tests for the Cambridge EFL examinations. UCLES' approach to ensuring these objectives can be met is based on a network of professionals with various levels of overlapping

responsibility, and on a set of procedures which applies to each professional level.

In the network of professionals there are three levels, in addition to UCLES' own staff. At the operational level are the OEs. At the next level up, in countries where there are sufficient numbers of OEs to merit it, Team Leaders (TLs) have responsibility for the professional supervision of OEs, in a ratio of about 1 TL to between 5 and 30 OEs. Finally, in countries where the number of TLs (and hence OEs) merit it, Senior Team Leaders supervise TLs in an average ratio of 1 STL :15 TLs.

The levels in this hierarchy are not sealed off from each other; it is a requirement that TLs and STLs must also be practising OEs in order to ensure that they can draw on their experience when it comes to dealing with the concerns of OEs. The hierarchical structure also enables a two-way channel of communication to be maintained up and down the levels.

The set of procedures which regulate the activities of these three professional levels is summarised by the acronym RITCME – Recruitment, Induction, Training, Co-ordination, Monitoring, Evaluation. Each procedure is defined by a list of Minimum Professional Requirements, which set down the minimum levels and standards (for recruitment, induction, training programmes, etc.) that must be achieved in order to meet the professional requirements of administering the Speaking Tests and sustain a fully effective Team Leader System.

3. Code of practice

In 1994 the members of ALTE adopted a formal Code of Practice, to make explicit the standards they aim to meet, and to acknowledge the obligations under which they operate. In doing this they highlighted the roles of those who have an interest in the setting and maintaining of standards in language examinations: test developers, test-users and test-takers.

Members of ALTE undertake to safeguard the rights of examination takers by striving to meet the standards of a Code of Practice in four areas:

- developing examinations
- interpreting results
- striving for fairness
- informing examination takers

Like examination developers, examination users – teachers, Directors of Studies, etc. – have a duty towards candidates, and are under an obligation to set and maintain high standards of fair behaviour. These responsibilities are described under four headings:

- selecting appropriate examinations
- interpreting results
- striving for fairness
- informing examination takers

A particular strength of this attempt to develop a code of practice is its acknowledgement that test development and use involve a shared responsibility between stakeholders.

4. Support and information for stakeholders

A major responsibility highlighted in the ALTE Code of Practice is that of providing adequate information to stakeholders.

UCLES provides Handbooks for each examination, giving detailed descriptions of test content, structure and assessment, together with sample question papers. Past paper packs, including listening cassettes and markschemes, are also available for each examination. For some examinations Examination Reports are produced annually containing useful information for teachers on the performance of candidates. UCLES has regular communication with publishers in the field who themselves produce books and other materials related to the Cambridge EFL examinations. UCLES also publishes regular newsletters and bulletins giving stakeholders information on a range of subjects – such as the CPE revision.

Issues of recognition and currency are important for all test-users, and UCLES publishes detailed information on recognition of the examinations by universities and colleges in the UK and North America, and is working to provide more comprehensive information on recognition by educational institutions and employers throughout the world.

While paper-based and electronic support are clearly valuable in helping us to fulfil our responsibility in disseminating information, face-to-face contact is often even more valuable and valued. UCLES holds a range of seminars about its tests, aiming to:

- offer information, support and updates to teachers who are already familiar with the tests
- present established tests to teachers who are new to the UCLES range
- introduce new tests

Investigating the impact of international language examinations

5. Special circumstances

Although tests are designed to assess language ability without being biased towards candidates from a particular culture or background, the standard format of the test is not always appropriate for candidates with certain disabilities or medical conditions. UCLES provides special Braille and large print versions of question papers for blind or partially-sighted candidates. Other special versions such as lip-reading versions of listening tests have been developed for hearing impaired candidates. Special arrangements can be made for candidates with other disabilities. For example, candidates with dyslexia or a physical difficulty can be given extra time to complete the test or an amanuensis to assist with writing.

There are also situations where it is not possible to arrange a paired speaking test – such as in a prison or a closed religious order; in these cases, special procedures are followed by the oral examiner to ensure that the candidate is treated in a fair and standardised way.

References and further reading:

Alderson, J C and Buck, G (1993): 'Standards in testing: a study of the practice of UK examination boards in EFL/ESL testing' *Language Testing*, 10/1,1-26

Hamp-Lyons, L (1997): 'Washback, impact and validity: ethical concerns' *Language Testing*, 14/3, 295-303

Rea-Dickins, P (1997): 'So, why do we need relationships with stakeholders in language testing? A view from the UK' *Language Testing*, 14/3, 304-314

Nick Saville, Manager, EFL Test Development and Validation Group, UCLES

Tests provided by major testing agencies and examination boards like UCLES have an impact on educational processes and on society in general. While washback, the effect of tests on language teaching and learning, has received some limited attention in recent years, the more general concept of impact and how it may be investigated has not been systematically addressed in this field. This paper discusses the concept of test impact in relation to international examinations such as those produced by UCLES EFL and reports on a long-term research programme which has been established to investigate the impact of the International English Language Testing System (IELTS).

IELTS is jointly owned by UCLES, the British Council and IDP Education Australia and is currently taken at 224 centres in 105 countries, by over 100,000 candidates per year – most of whom are seeking admission to higher education or training in the UK, Australia, Canada and the USA (for details of test format see the IELTS Handbook).

Following the most recent revision of IELTS in 1995, it was decided that the impact of IELTS should be investigated systematically. In a working paper in 1996 entitled 'Considering the Impact of Cambridge EFL Examinations,' Milanovic and Saville noted

> *the complex interactions between the factors which make up the teaching and learning context surrounding the Cambridge examinations – including the individual learners, the teachers, the classroom environment, the choice and use of materials and so on.*

The same complex interactions exist for IELTS as for other Cambridge examinations (although IELTS was not originally designed to have the same curriculum impact as examinations such as FCE, CAE or CPE). For example, there are at least ten published textbooks with the IELTS name in their titles which are now widely used. In 1996, therefore, a project was set up to begin looking at ways in which the impact of IELTS could be investigated more effectively. This was co-ordinated by Nick Saville and Mike Milanovic at UCLES working in conjunction with Charles Alderson at Lancaster University who was commissioned to help develop the Impact Study.

Stakeholders in language testing

The factors which make up the teaching and learning context surrounding an examination like IELTS impact on a range of stakeholders involved with the examination. These stakeholders form the testing community in which we, as an international examination board, are located – what might be termed our language testing constituency (see the previous article by Lynda Taylor for further discussion). Relationships between the stakeholders entail certain roles and responsibilities, just as in any other community, and the specific work which is going on as part of the IELTS project is designed to help us understand the impact of the testing system within the wider IELTS constituency.

Having identified these stakeholders, what does it then mean for an examination board like UCLES, to be accountable in its relationships with these other stakeholders? In practice, it means that we must seek to involve as many stakeholders as possible and be prepared to review and change what we do in light of what we find out about how they use the examinations and what they think about them.

The IELTS Impact projects

The IELTS Impact projects can be seen as an extension of the consultation and piloting activities, discussed by Lynda Taylor in the previous article (p 2). In 1995 IELTS was introduced in its latest revised form, and at that time, it was agreed that procedures would be developed to monitor the impact of the test and to contribute to the next revision cycle – starting around 2000 or 2001. In order to understand the test impact better and to conduct effective surveys to monitor it, it was decided that a range of standardised instruments and procedures should be developed to focus on the following aspects of the test:

- the content and nature of classroom activity in IELTS-related classes
- the content and nature of IELTS teaching materials (including textbooks)
- the views and attitudes of user groups towards IELTS
- the IELTS test-taking population and the use of test results

The first two of these points concern washback in the traditional sense – the effect of the test on teaching and learning. The second two are concerned with the wider impact of the test in terms of the effects of the test on other systems in the administrative and academic contexts where the tests are used, and on attitudes and behaviour of the stakeholders.

This long-term study comprises three phases as shown in Table 1.

Phase 1	**1995-1996**	**The identification of areas to be targeted and the development of instrumentation to collect information which allows impact to be measured**
Phase 2	**1997-1999**	**The validation of the instruments prior to full-scale implementation**
Phase 3	**2000**	**Implementation of the instruments as part of a major survey**

Table 1

We are currently at the end of Phase 2 and are looking to implement the instruments in Phase 3 during 2000 as part of a major survey to review the impact of the current test and to help formulate the scope of the next revision project.

The lack of appropriate validation of such instruments has been noted, for example by Alderson and Banerjee (1996). In all cases, therefore, the aim has been to validate the instruments in Phase 2 before proceeding with large-scale data collection. An approach to validation borrowed from language testing has been applied to the instruments developed for these projects. This approach sets out to establish the validity, reliability and practicality of the instruments using both quantitative and qualitative methods.

The initial development work was documented by researchers at Lancaster (Banerjee 1996, Herrington 1996, Horak 1996, Winetroube 1997). It is intended that the different projects will eventually be written up as Working Papers in the area of Impact and that these will be combined into a volume in the UCLES/CUP Studies in Language Testing Series. This volume will cover both the development and validation Phases of the project.

Four sub-projects focus on areas of crucial importance in considering the impact of IELTS and each sub-project has associated instrumentation (cf. IELTS Annual Review 1998/9).

Project One: the context and nature of classroom activity in IELTS classes

Four instruments and associated procedures were developed:

1. an observation schedule for classroom activity
2. a procedure for producing summaries of classroom activity
3. a questionnaire for teachers after teaching an observed lesson
4. a questionnaire for students after taking part in an observed lesson

The validation of these instruments is still in progress. Some of the early materials were trialled on a small-scale during the development phase by staff and students at Lancaster University but subsequently this project has been difficult to implement because of the requirement for support at a local level within teaching establishments.

Recently we have been seeking to set up more extensive trials by contacting individuals who have a research interest in this area. It is hoped that Phase 2 for this project will be completed in 2000. In particular this links up with a research project funded by the IELTS Research Programme to be conducted by John Read and Belinda Hayes entitled 'The Impact of the IELTS Test on Preparation for Academic Study in New Zealand'.

Project Two: the content and nature of IELTS teaching materials

An instrument was developed to capture information about IELTS preparation material. This was based on a number of sources including the ALTE checklists and commissioned work carried out by post-graduate students at Lancaster University. In Phase 2 the checklists were validated by conducting trials on a range of IELTS and other materials using three raters. In addition a survey of textbook use was carried out as well as interviews with writers of IELTS books.

To conclude Phase 2, an interim report, an evaluation and a major revision of the checklists have recently been compiled by UCLES EFL with Dr Roger Hawkey providing external consultancy. It is intended that this will be a published Working Paper and it has already provided input to conference presentations, for example the presentations by Nick Saville at TESOL Vancouver in March 2000 and in Dubai in May 2000.

Later this year further surveys of textbook use and the reapplication of the instruments to IELTS and other examination materials will be carried out. The results of this will be reported later this year as part of the Phase 3 work.

Project Three: the views and attitudes towards IELTS user groups

Seven instruments were developed to explore the views and attitudes of a wide population of IELTS users:

1. a questionnaire for students preparing for IELTS
2. a questionnaire for teachers preparing students for IELTS
3. a questionnaire for teachers preparing students for academic study (post-IELTS)
4. a questionnaire for IELTS administrators
5. a questionnaire for admissions officers in receiving institutions
6. a questionnaire for students who have taken IELTS
7. a questionnaire for academic subject teachers

In Phase 2 some of these instruments were administered to IELTS stakeholders. As a result of analysis (both statistical and qualitative) which formed part of a workshop at UCLES conducted by Dr Antony Kunnan (Spring 1999), it was decided to rework the instruments before proceeding with additional data collection. The revision and report on this area was commissioned by UCLES and was undertaken by Dr Kunnan. The final report and revised questionnaires were submitted at the end of 1999.

For Phase 3, the revised questionnaires will be used to survey a wide-range of IELTS stakeholders and the results will be compiled for a report by end-2000. Collaboration from the IELTS partners and a range of IELTS stakeholders will be required if this is to be successful.

Project Four: the IELTS test-taking population

To supplement the routine information collected about the IELTS candidature a detailed Candidate Information Sheet was developed. This instrument – the in-depth Candidate Information Sheet – focusing on IELTS candidates was adapted from a range of existing questionnaires for learner profiling (such as the UCLES/UCLA Language Learning Questionnaires – Bachman et al). It includes traits focusing on attitude, motivation and cognitive/metacognitive features, as well as additional standard demographic data.

In Phase 2 this questionnaire was administered to a wide range of IELTS candidates and as a result of work carried out by UCLES validation staff working with Dr Jim Purpura, a revised instrument has now been developed. Purpura has documented the use of Structural Equation Modelling (SEM) as a method of validating questionnaires and this approach was applied to the IELTS instrument (see Purpura 1999).

In Phase 3, the revised instrument will be administered again and the responses from candidates will be linked to their performance on the test items. A report will be produced by end of this year.

Summary

As a result of the work carried out in Phase 3 during 2000, it is hoped that it will be possible to reach some conclusions about the impact of IELTS on our stakeholder community which can be substantiated by the extensive research set out above.

It is hoped that this work will help to demonstrate that IELTS, along with other Cambridge EFL examinations, has positive educational impact and that the research itself will make a positive contribution to the field in this area. Progress on Phase 3 of the project will be reported in future editions of Research Notes.

References and further reading

Alderson, J C and Banerjee, J (1996): 'How might Impact study instruments be validated?' A paper commissioned by the University of Cambridge Local Examinations Syndicate as part of the IELTS Impact Study

Elder, C (ed) (2000): *Experimenting with uncertainty*, Studies in Language Testing, Volume 11, Cambridge University Press

Alderson, J and Wall, D (1993): 'Does washback exist?' *Applied Linguistics* 14/2, 115-129

Association of Language Testers in Europe (1998): *ALTE Handbook of Language Examinations and Examination Systems*, University of Cambridge Local Examinations Syndicate

Bachman, L (1990): *Fundamental considerations in language testing*, Oxford University Press

Bachman, L and Palmer, A (1996): *Language Testing in Practice: Designing and Developing Useful Language Tests*, Oxford University Press

Bailey, K (1996): 'Working for washback: a review of the washback concept in language testing' *Language Testing*, 13/3, 257-279

Banerjee, J (1996): *The Design of the Classroom Observation Instruments*, UCLES Internal Report, University of Cambridge Local Examinations Syndicate

Cheng, L (1997): 'How does washback influence teaching? Implications for Hong Kong,' *Language in Education* 11/1, 38-54

Hamp-Lyons, L (1997): 'Washback, impact and validity: Ethical concerns,' *Language Testing* 14/3, 295-303

Herrington, R (1996): *Test-taking strategies and second language proficiency: Is there a relationship?* Unpublished MA dissertation. Lancaster University

Horak, T (1996): *IELTS Impact Study Project*, Unpublished MA assignment, Lancaster University

Macnamara, T (1999): 'Validity in Language Testing: The Challenge of Sam Messick's Legacy,' Messick Memorial Lecture, *Language Testing Research Colloquium*, Japan

Milanovic, M and Saville, N (1996): *Considering the Impact of Cambridge EFL Examinations*, UCLES Internal Report, University of Cambridge Local Examinations Syndicate

Purpura, J (1999): *Learner strategy use and performance on language tests,* Studies in Language Testing Volume 8, Cambridge University Press

Saville, N (1998): *Predicting impact on language learning and the classroom*, Internal draft for IELTS Impact Study, University of Cambridge Local Examinations Syndicate

Shohamy, E (1999): 'Language Testing: Impact' in Spolsky, B (ed) (1999): *Concise Encyclopedia of Educational Linguistics,* Pergamon

Taylor, L (1999) 'Constituency Matters,' Paper presented at the Language Testing Forum, Edinburgh, November 1999

The IELTS Handbook (1999): University of Cambridge Local Examinations Syndicate

The IELTS Annual Review 1997/8 (1998): University of Cambridge Local Examinations Syndicate, The British Council, IDP Education Australia

IELTS Specimen Materials Handbook, and Specimen Materials (1997): University of Cambridge Local Examinations Syndicate, The British Council, IDP Education Australia

Winetroube, S (1997): *The design of the teachers' attitude questionnaires*, UCLES Internal Report, University of Cambridge Local Examinations Syndicate

Yue, W (1997): *An investigation of textbook materials designed to prepare students for the IELTS test: a study of washback*, Unpublished MA dissertation, Lancaster University

The UCLES EFL item banking system

Simon Beeston, EFL Validation Manager, UCLES

What is an item bank?

An item bank is, typically, a large collection of test items which have been classified and stored in a database so that at a later date, they can be selected for use in new tests. The items are all classified according to certain descriptive characteristics such as the topic of a text, the testing point for an item, etc., as well as statistical information about how difficult each item is. It is important that all of the item difficulties have been located on a common scale of difficulty so that any combination of items can be put into a new test and the item difficulties added together to give a precise measure of the difficulty of that test. When the test items have been selected, the test paper itself needs to be produced, so that a desktop publishing package can be used to prepare the test for printing.

It is only in the last few years that UCLES EFL have been using complex software to manage and develop our tests. What we have done to develop the standard methodology outlined above is to take our existing test production procedures and see how best to fit them into an item banking paradigm. One of the key features of Cambridge EFL examinations is the emphasis on quality control so it has been important to make sure that the key stages of test development procedures are captured by the software used.

Quality control

Many organisations now endeavour to practise Total Quality Management by which they mean that they adopt a comprehensive approach to achieving quality in every aspect of their work. For UCLES EFL, this starts with ensuring that we know all about the different kinds of people who take our examinations and exactly what it is they need and expect when they enter an examination. Not surprisingly, we have identified issues of fairness and the usefulness of our qualifications as key requirements of our examinations. Part of what fairness in language testing means is making sure that procedures for every stage in the testing process are well planned and carefully managed, including the way each test is produced, and the way it is administered, marked and graded. Our approach to item banking addresses the way the test is produced by guaranteeing that all test material goes through a series of specific quality control checks before a completed test is constructed and administered.

Reviewing new material

This process starts with the development of detailed test specifications which provide precise guidelines for the production of an examination. These specifications are used by the UCLES Subject Officer and the Chair of the item writer team when reviewing prospective material to determine, for example, whether the topic and discourse features of a text are suitable and if it will be able to support the required number of questions. If material is accepted at this meeting, it is returned to the item writers who can then complete their work on the material by writing the required number of questions. The material is then submitted to UCLES again for an editing meeting. At the editing meeting, the Subject Officer, Chair of the item writer team and the team of item writers, review all of the submitted material and decide if it is of suitable quality. This involves determining whether the questions are adequately measuring the test construct. That is to say, when editing material for a test of reading, for example, it should not be possible to guess the answers on the basis of logical deduction or background knowledge of the subject. Where examples are given to guide the candidate, it must be clear that the processes involved when answering the questions are the kinds of processes we typically associate with reading. In summary, the editing meeting is one of the stages that help ensure our tests are fair for all candidates.

The start of the item banking process

Significantly, all of this takes place before any material is entered onto the item banking system. In fact, it is only material which is accepted at the editing stage that then goes forward to be word processed and stored on the system. The item banking database currently has over 100 different banks of test material, comprising approximately 90,000 items and this number is expected to rise to around 100,000 in the next few months. These items are usually grouped together as tasks such as a reading test with 10 items (questions) or a gapped text with 15 items. Every examination has a number of banks holding its material. These banks correspond to the key stages of test development. Accordingly these banks are called the Edited, Pretest, Test Construction and the Live banks.

The Edited bank

Each task is described using a set of attributes (database fields) which classify the task according to the text type, topic, source of the material and so on. Typically each task uses around 10 to 15 attributes with Listening tasks using slightly more as the accent and age of the speaker are also recorded as attributes. It is not possible for staff to access the system unless they have a password and this only allows them into the banks on which they work. The database is also encrypted making it extremely difficult for anyone to see material unless they are an authorised member of staff. The system runs in conjunction with Microsoft Word so that all tasks are stored as fully word-processed documents. These documents are stored within the encrypted database and inaccessible to anyone but a designated item bank user. As soon as the material has been entered onto the system and attributes added, the new tasks are printed off and sent to the Chair of the item writer team to check for any errors. On return of the material, any errors in the tasks are immediately rectified on the database.

Pretesting

At this point, the tasks are combined into pretests and moved into the Pretest bank. The Pretests are created automatically by the system which copies and pastes all of the required tasks and associated files (front pages, examples, etc.) into a Microsoft Word document where the part, rubric (instructions) and item numbers are all automatically created. The Pretest is printed and checked by the Subject Officer who then passes it to the Pretesting Unit. The Pretesting Administrator arranges the pretesting of approximately 200 separate pretests each year involving some 30,000 candidates around the world. All of this takes place using centres who have undertaken to treat the pretests as they would live material even to the extent of returning it to Cambridge by secure courier delivery services. Once in Cambridge, the answer sheets are marked and answer keys amended and developed in the light of this process. The answer sheets are scanned and scores returned to each centre thus providing not just examination practice but feedback on how well each candidate performed. The resulting data are analysed and filed along with comments from teachers and candidates. These files are sent to the Subject Officer and Chair of the item writer team who then meet with the item writers to review the material in the light of the statistical information and centre/candidate feedback. The Pretest Review meeting is therefore the next quality control stage where the pretests are evaluated on the basis of whether the individual items discriminated between the stronger and weaker candidates and if the items were of the appropriate level of difficulty. It is at this stage that the soft feedback is reviewed to ensure that there were no problems in terms of the suitability of the material. At this stage, material can be rejected if unsuitable, returned to the Edited bank to be rewritten and pretested again, or passed onto the Test Construction bank.

Before tasks are moved out of the Pretest bank, all statistical information is loaded into the database. In addition to classical item statistics (facility and discrimination), item difficulties are also estimated and loaded into the item bank. These item difficulties are derived using a type of analysis called Rasch analysis which relates the items to each other on the basis of common items in different pretests (anchor items). The item difficulties are therefore anchored to a common scale thus making it possible to recombine tasks in the item bank but still add up the specific item difficulties to find out how difficult a test will be for its intended candidature.

The Test Construction bank

Tasks are moved into the Test Construction bank and printed off for further proofing by the Chair of the item writer team. Once again, any editing changes are immediately made to the tasks in the Test Construction bank. The bank is now ready for live tests to be constructed. This is done on-line when the Subject Officer and the Chair of the item writer team meet and work on a networked PC to select the combination of tasks and items they wish to include in the new test. For any possible selection, the item difficulties can be immediately calculated by the system to predict the overall difficulty of the test. This allows new versions to be constructed that are equivalent to previous versions and thus address a fundamental issue in testing; that of fairness. Once again, a report listing all of the attributes describes the content of the test and allows the test constructors to determine whether the test adequately covers the required range of testing points. Following this meeting and the automatic creation of the test, there is an Examination Ratification meeting where all of the papers for a particular administration are looked at together to consider the overall coverage of the whole examination. If no changes are required, the test in question is moved to the Live bank where it begins a process of Question Paper Production (QPP).

The Live bank

Once in the Live bank, a default schedule is added to the test which automatically calculates the dates by which each event in the schedule needs to be completed by if the test is to be ready on time. These events include stages such as further vetting, proofing, typesetting and finally printing. An average QPP schedule has around 15 quality control stages which need to be signed off in the database. If an event is not checked as complete, the item bank server sends an automatic e-mail to the Group Manager responsible for the examination, alerting them to the problem. Finally, the test is printed and despatched to centres.

Validity and reliability

It would not be possible to conclude this brief account of the UCLES EFL test production methodology without saying something about how this process addresses issues of validity and reliability. Validation is often described as the process of building an argument to support the inferences that are made from test scores. For Cambridge EFL examinations, that process is greatly assisted by the systematic review of new items throughout the stages identified above. Similarly, the pretesting stage identifies items which may be performing poorly for some reason. By removing these items, the remaining material is of a better quality and will measure more reliably. Although much validation work is carried out post hoc after the test administration using live data, the effort that goes into producing the tests can clearly be seen as contributing to the overall validity, reliability and, above all, the quality of Cambridge EFL examinations.

Development of new item-based tests: The gapped sentences in the revised CPE Paper 3

David Booth, Nick Saville, EFL Test Development and Validation Group, UCLES

The development of new test items fits into a general model of test development which UCLES has established. The model is essentially cyclical and involves incremental stages of development and evaluation (see Cambridge First 6, June 1999). The development of the gapped sentence items for the revised Cambridge Proficiency in English (CPE) Paper 3 is an example of the implementation of this model of test development.

The development of new testing techniques often relies on developments in the field of applied linguistics. One area which has become a focus of academic enquiry in recent years has been a renewed interest in lexis. This new interest has stemmed from two main areas. Firstly, theoretical work on the acquisition of language, in particular the acquisition of lexicalised chunks of language. This work was initially inspired by the work of Pawley and Syder (1983). More recently researchers working in the area of phraseology have started to explore a model of collocation and assess how the use of appropriate collocations may relate to linguistic sophistication – which in turn is associated with high level language proficiency. In particular the work of Cowie and Howarth (1996) and Howarth (1999) has helped develop our understanding of lexical units and collocation. Alongside this theoretical approach has been a great increase in observable data on lexis available from corpora of English, such as the Cobuild Corpus, the British National Corpus (BNC) and the Cambridge International Corpus (CIC). These corpora and their search engines have provided linguists and lexicographers with new insights into lexical combinations and the use of collocation.

The increased data available in general English corpora, along with the development of the Cambridge Learner Corpus in written English has led to renewed interest in the testing of collocation at UCLES, particularly in the context of the CPE Revision Project (Research Notes March 2000).

CPE is the highest examination offered by UCLES EFL. Work by Cowie and Howarth (1996) has indicated that greater use of collocations produces more sophisticated language. For this reason, items focusing on collocation were felt to be particularly appropriate at CPE level.

A thorough and informative article on this area of work has been written by Peter Hargreaves, Director UCLES EFL (Hargreaves 2000). This work attempts to identify the type of collocations which could be tested at an advanced level, such as CPE level at ALTE Level 5.

To illustrate here is an example taken from Hargreaves (2000) of a possible collocation item type:

Circle the word from among A – D which fits most appropriately in the blanks in all three sentences (i-iii):

A fashion B opinion C feeling D will

i You cannot simply come into an existing situation and impose your _____ on everyone like that

ii Though he may have good reasons for introducing such measures, popular _____ is likely to prevent them from working

iii She may insist on such a dress code in the office, but whether it's correct to do so is a matter of _____

The initial development of the collocation items were within the context of CPE Paper 1. Traditionally this paper, although a reading paper, had also focused on the vocabulary resource of candidates at this level. The items were written by very experienced UCLES item writers based on the multiple choice format of the paper. After trialling and evaluation, however, it was found that such items were not suited to the multiple choice format as it was difficult to write clear distractors.

The evaluation concluded that the tasks would be better as productive tasks, where candidates provided the answer rather than chose the answer from alternatives.

This led to the revised items being developed for CPE Paper 3 – Use of English – which has a grammatico-lexical focus and includes productive items.

An example of this type of item, again from Hargreaves (2000) is below:

The sentences below can be completed appropriately using the same single word in each blank. Write what you think this word is in the box provided.

i I have _____ faith in your judgement

ii The _____ insult in his speech was clear to most of the journalists present

iii He failed to see the potential difficulties _____ in such an aggressive management style

The validation of these items involved more cycles of development and evaluation. Input was sought from experts in the field of collocation, for example Howarth, and items were subjected to trialling and statistical analysis using classical test analysis and IRT. Experienced freelance test writers were also commissioned to assess the viability of consistently producing these 'gapped sentences' at the appropriate level of difficulty.

The most recent data on the gapped sentence item types indicate that they are appropriate for the CPE or mastery level candidate. They are challenging and impact positively on the candidate's learning of lexical meanings. They also reflect an aspect of linguistic competence which has not always been clearly defined or effectively tested.

Background to the validation of the ALTE 'Can-do' project and the revised Common European Framework

References and further reading

Association of Language Testers in Europe (1998): *ALTE Handbook of Language Examinations and Examination Systems*, University of Cambridge Local Examinations Syndicate

Cambridge First 6, June 1999, p.2, University of Cambridge Local Examinations Syndicate

Cowie, A and Howarth, P (1996): 'Phraseological Competence and Written Proficiency' in Blue, G (ed) *Language and Education,* Multilingual Matters

Hargreaves, P (2000): 'How Important is Collocation in Testing the Learners Language Proficiency?' in Lewis, M (ed) *Teaching Collocation – further developments in the Lexical Approach,* Language Teaching Publications

Howarth, P (1999): 'Phraseology and Second Language Proficiency', *Applied Linguistics* 19/1

Pawley, A and Syder, F H (1983): 'Two puzzles for linguistic theory: nativelike selection and nativelike fluency' in Richards, J C and Schmidt, R W (eds) *Language and Communication,* Longman

Neil Jones, Grading Co-ordinator, UCLES

This article includes excerpts from the Appendix to the Council of Europe Framework document due to be published in 2001.

The ALTE Framework and the Can-do project

The ALTE Framework

The ALTE Can-do statements constitute a central part of a long-term research programme set by ALTE, the aim of which is to establish a framework of critical levels of language performance, within which examinations can be objectively described.

Much work has already been done to locate the examination systems of ALTE members in this framework, based on an analysis of examination content and task types, and candidate profiles. A comprehensive introduction to these examination systems is available in the ALTE Handbook of European Language Examinations and Examination Systems.

The ALTE Can-dos are user-orientated scales

The aim of the Can-do project is to develop and validate a set of performance-related scales, describing what learners can actually do in the foreign language.

In terms of Alderson's (1991) distinction between constructor, assessor and user-orientated scales, the ALTE Can-do statements in their original conception are user-orientated. They assist communication between stakeholders in the testing process, and in particular the interpretation of test results by non-specialists. As such they provide:

a) a useful tool for those involved in teaching and testing language students. They can be used as a checklist of what language users can do and thus define the stage they are at

b) a basis for developing diagnostic test tasks, activity-based curricula and teaching materials

c) a means of carrying out an activity-based linguistic audit, of use to people concerned with language training in companies

d) a means of comparing the objectives of courses and materials in different languages but existing in the same context

They will be of use to people in training and personnel management, as they provide easily understandable descriptions of performance, which can be used in specifying requirements to language trainers, formulating job descriptions and specifying language requirements for new posts.

The ALTE Can-dos are multilingual

An important aspect of the Can-dos is that they are multilingual, having been translated so far into twelve of the languages represented in ALTE. These languages are: Catalan, Danish, Dutch, English, Finnish, French, German, Italian, Norwegian, Portuguese, Spanish and Swedish. As language-neutral descriptions of levels of language proficiency they constitute a frame of reference to which different language examinations at different levels can potentially be related. They offer the chance to demonstrate equivalences between the examination systems of ALTE members, in meaningful terms relating to the real-world language skills of people achieving a pass in these examinations.

Organisation of the Can-do statements

The Can-do scales consist currently of about 400 statements, organised into three general areas: Social and Tourist, Work, and Study. These are the three main areas of interest for most language learners. Each includes a number of more particular areas, e.g. the Social and Tourist area has sections on Shopping, Eating out, Accommodation etc. Each of these includes up to three scales, for the skills of Listening/Speaking, Reading and Writing.

Each such scale includes statements covering a range of levels. Some scales cover only a part of the proficiency range, as of course there are many situations of use which require only basic proficiency to deal with successfully.

Assumptions of the Can-do approach

Levels describe typical patterns of ability

The Can-do scales have been subjected to an extended process of empirical validation. The validation process is aimed at transforming the Can-do statements from an essentially subjective set of level descriptions into a calibrated measuring instrument. This is a long-term, ongoing process, which will continue as more data become available across the range of languages represented by ALTE.

So far data collection has been based chiefly on self-report, the Can-do scales being presented to respondents as a set of linked questionnaires. Nearly ten thousand respondents have completed questionnaires. For many of these respondents, additional data are available in the form of language examination results. This is believed to be by far the biggest collection of data ever undertaken to validate a descriptive language proficiency scale.

Thus it is the typical response patterns of this large sample of respondents which define the meaning of a given level in can-do terms. In other words, the definition of a level is not based on a priori prescriptive, absolute criteria, but is rather descriptive of the experience of a large number of foreign language users.

So far respondents have been predominantly European language speakers, and it is likely therefore that the Can-do scales reflect European patterns of foreign language skills. The levels describe profiles of language skill which are typical for Europeans, and speakers of European languages – in terms of their relative abilities in reading, writing or face-to-face communication, for example. This makes for ease of use, because in a European context (and probably in many other contexts) the simple level classification constitutes a rich description of a learner's probable skills profile. Of course, this does not preclude a more analytic use of the levels, so that an individual could be described as, say, 'Level 4 generally but only Level 2 in writing.'

Respondents should be matched to appropriate questionnaires

Questionnaires have been administered in the subjects' own first language, except at very advanced levels. Respondents have been matched to appropriate questionnaires – the Work scales given to people using a foreign language professionally, the Study scales to respondents engaged in a course of study through the medium of a foreign language, or preparing to do so. The Social and Tourist scales are given to other respondents, while selected scales from this area have also been included in the Work and Study questionnaires as an 'anchor'.

The systematic use of anchor statements is necessary to enable the relative difficulty of the areas of use, and particular scales, to be established. The use of Social and Tourist scales as an anchor was based on the assumption that these areas call upon a common core of language proficiency and can be expected to provide the best point of reference for equating the Work and Study scales.

Can-do scales are language neutral

A rather fundamental assumption is that it is possible to construct Can-do descriptions of language level which are valid irrespective of the language background of the learner or the target language being studied.

However, it is possible to imagine that speakers from particular linguistic or cultural groups may experience particular language tasks as being more or less difficult – paying compliments, complaining about service, or making small-talk, for example. The data-based validation of the Can-dos allows this interesting question to be investigated.

Relating the Can-dos to ALTE examinations

Following the initial calibration of the Can-do statements, and the textual revision described above, attention has turned to establishing the link between the Can-do scales and other indicators of language level. In particular we have started looking at performance in ALTE

examinations, and to the relation between the Can-do scales and the Council of Europe Framework levels.

Beginning in December 1998, data were collected to link Can-do self-ratings to grades achieved in Cambridge EFL examinations at different levels. A very clear relationship was found, making it possible to begin to describe the meaning of an examination grade in terms of typical profiles of Can-do ability.

A conceptual problem to be addressed in this context concerns the notion of mastery – that is, what exactly do we mean by 'can do'? A definition is required in terms of how likely we expect it to be that a person at a certain level can succeed at certain tasks. Should it be certain that the person will always succeed perfectly on the task? This would be too stringent a requirement. On the other hand, a 50 per cent chance of succeeding would be too low to count as mastery.

The figure of 80 per cent has been chosen, as an 80 per cent score is frequently used in domain or criterion-referenced testing as an indication of mastery in a given domain. Thus, candidates achieving an ordinary pass in an ALTE examination at a given level should have an 80 per cent chance of succeeding on tasks identified as describing that level.

By defining 'can do' explicitly in this way we have a basis for interpreting particular ALTE levels in terms of Can-do skills.

While the relation to examination performance has so far been based on Cambridge EFL examinations, data linking Can-dos to performance in other ALTE examinations will continue to be collected, allowing us to verify that these different examination systems relate in essentially the same way to the ALTE 5-level framework.

Anchoring to the Council of Europe Framework

In 1999 responses were collected in which anchors were provided by statements taken from the 1996 Council of Europe Framework document. Anchors included:

1. the grid of major categories of language use by level identified as 'Table 7' in Council of Europe (1998), 133
2. 16 statements from scales relating to spoken interaction (Fluency)

Table 7 was chosen because in practice it is achieving wide use as a summary description of levels. ALTE's ability to collect response data in a large number of languages and countries provided an opportunity to contribute to the validation of the scales in Table 7.

The Fluency statements had been recommended because they had been found to have the most stable difficulty estimates when measured in different contexts in the Swiss project (North 1996/2000). It was expected that they should thus enable a good equating of the ALTE Can-dos to the Council of Europe Framework.

Levels of proficiency in the ALTE Framework

At the time of writing the ALTE Framework is a five-level system. The validation described above confirms that these correspond broadly to levels A2 to C2 of the CE Framework. Work on defining a further initial level (Breakthrough) is in progress, and the Can-do project is contributing to the characterisation of this level. Thus the relation of the two frameworks can be seen as follows:

Council of Europe	A1	A2	B1	B2	C1	C2
ALTE	Breakthrough	1	2	3	4	5

The salient features of each ALTE level are as follows:

ALTE Breakthrough Level: a basic ability to communicate and exchange information.

ALTE Level 1 (Waystage User): people are able to deal with simple, straightforward information and begin to express themselves in familiar contexts.

ALTE Level 2 (Threshold User): in familiar situations, users can express themselves in a limited way and deal in a general way with non-routine information.

ALTE Level 3 (Independent User): the salient feature is instrumental, functional ability – people can achieve most goals, and express themselves on a range of topics.

ALTE Level 4 (Competent User): the salient feature is how well people can do it, in terms of appropriacy, sensitivity, and the capacity to deal with unfamiliar topics.

ALTE Level 5 (Good User): moves beyond purely instrumental ability (that is, the capacity to get things done). The salient feature is linguistic. It indicates a capacity to deal with material which is academic or cognitively demanding, and to use language to good effect. That is, it describes a level of performance which may in certain respects be more advanced than that of an average native speaker.

References and further reading

Alderson, J C (1991): 'Bands and scores' in: Alderson, J C and North, B (eds.): *Language testing in the 1990s*, British Council/Macmillan, Developments in ELT, 71–86

Council of Europe (1998): *Modern Languages: Learning, Teaching, Assessment. A Common European Framework of Reference*, rev IV, Council of Europe

North, B (1996/2000): *The development of a common framework scale of language proficiency*, PhD thesis, Thames Valley University, Peter Lang

Investigating the paired speaking test format

Lynda Taylor, Performance Testing Co-ordinator, UCLES

A face-to-face speaking test has been an integral component of Cambridge EFL examinations since shortly after the Certificate of Proficiency (CPE) was first introduced. Until fairly recently, the traditional or standard format for most speaking tests has been the singleton or one-to-one format, i.e. one candidate in an oral interview with one examiner. In the early 1980s, however, UCLES EFL explored the use of a paired (two candidates and two examiners) and a group (three candidates and two examiners) test format; these alternative formats were offered as options alongside the traditional single-candidate format for established examinations such CPE and FCE, as well as for newer tests such as the Preliminary English Test (PET).

The move towards using a paired (or group) format for assessing speaking ability directly reflected changes which were taking place during the 1980s in the teaching and learning of English as Foreign Language. Developments in applied linguistics during the 1970s had led to a better understanding of the communicative role of language and this in turn influenced approaches to language teaching; the focus shifted away from the teaching of knowledge about language towards developing the ability to use language for communicative purposes. In the classroom context, for example, this meant greater use of pairwork interaction and group discussion. The UK educational tradition has always been characterised by a close relationship between teaching and testing; it is not surprising, therefore, that developments in EFL pedagogy were in time reflected in changing approaches to assessment.

A paired candidate format was also a design feature of the Certificates in the Use of English as a Foreign Language (CUEFL), developed in the early 1980s by the Royal Society of Arts Examinations Board. In 1988 production and administration of the CUEFL transferred to UCLES and in 1989 this suite of examinations was revised to produce the Certificates in Communicative Skills in English (CCSE). Although a number of test features were changed, the paired candidate format for assessing speaking was retained.

The Certificate in Advanced English (CAE), introduced in 1991, included a speaking test in which the paired candidate and paired examiner format was obligatory. This format was preferred because it allowed for a more varied sample of interaction (candidate-candidate as well as candidate-examiner); it also provided for two assessments for each candidate (one from the interlocutor and one from the assessor), thus contributing to the reliability of the speaking test.

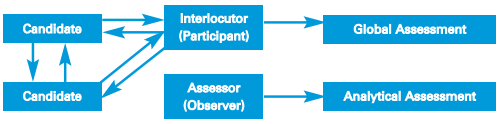

Figure 1: The three-way potential for interaction and the dual perspective of two raters in the paired Speaking Test format.

The paired candidate format was subsequently introduced as standard for the Key English Test (KET) in 1993, for revised PET in 1995 and for revised FCE in 1996.

The decision to adopt the paired candidate format as standard in many of the Cambridge EFL speaking tests has been based not only upon pedagogical considerations, but also upon the findings of various studies of spoken language discourse over the past decade; such studies highlight the extent to which discourse is influenced by its goals and participants.

Research into speaker interaction in the oral interview

Hughes (1989) drew attention to 'at least one potentially serious drawback' of the traditional interview format: the power relationship which exists between tester and candidate. He also suggested that 'only one style of speech is elicited and many functions … are not represented in the candidate's performance', adding that 'discussions between candidates can be a valuable source of information' (pp 104-108). Ross and Berwick's study (1992) showed how oral interviewers use features of control (e.g. topic nomination) and accommodation (e.g. speech modifications) for different purposes. Young and Milanovic's analysis of the one-to-one FCE interview (1992) indicated that the resulting examiner-candidate discourse was highly asymmetrical in terms of features of dominance, contingency and goal-orientation.

The fixed role relationship (examiner-candidate) in a one-to-one test format makes it difficult for the candidate to escape this asymmetry. The paired candidate format, on the other hand, provides the potential for various interaction patterns: between candidate and examiner; between the 2 candidates; and between the three participants. The asymmetrical nature of the discourse is therefore considerably reduced.

Within the context of the current CPE Revision, we have recently been analysing and comparing the quantity and quality of candidate language generated by the one-to-one and paired formats; the findings from these studies provide yet further support for the choice of the paired format as the preferred or standard approach.

A quantitative comparison of the paired and one-to-one format

One might expect the quantity of language produced by one paired candidate (as measured by the number of words and turns) to be lower overall than the amount produced by one single candidate in the CPE speaking test. Taking relative timings into account, it would be reasonable to assume that each candidate in the paired format would produce about 60% of the quantity of language produced by a single candidate; in reality, however, transcription study showed that the volume of language produced by each candidate in the pair was considerably larger than that, around 75% of the volume produced by a

single candidate. Furthermore, the overall contribution of the examiner (in terms of number of words and turns) was reduced in the paired format and the relative contribution of the two candidates increased. This suggests that a more balanced interaction is taking place between the participants in the discourse, with the examiner assuming a less dominant role, particularly during the long turn and during the final discussion phase; the problem of asymmetry highlighted by earlier studies is apparently reduced. In addition, the substantial increase in the number of turns from paired candidates and the significant variation in turn length across the different parts of the test suggest that the paired format is capable of generating a richer and more varied sample of spoken language from each candidate than is usually produced with the one-to-one format. In other words, the paired format appears to encourage greater interactiveness.

A qualitative comparison of the paired and one-to-one format

The findings from the quantitative study described above were complemented by results from a second study with a more qualitative focus. Building on the work of Bygate (1988) and Weir (1993), a recent survey of the literature on speaking ability identified a total of 30 communicative language functions which characterise spoken discourse; these can be broadly categorised as informational (e.g. expressing opinions), interactional (e.g. persuading), or to do with managing interaction (e.g. terminating a discussion). This list of functions has recently been used within UCLES EFL to develop observation checklists for the a priori and a posteriori analysis of task output in speaking tests (see Nick Saville's article, 'Using observation checklists to validate speaking-test tasks', p 16). The use of observation checklists provides a valuable complementary methodology to the more labour-intensive transcription methodology for test-task validation.

An a priori analysis of the one-to-one speaking test format for CPE suggested that 20 of the 30 functions could be expected to appear in a candidate's spoken output; a paired format, however, appeared capable of eliciting 28 of the 30 functions – a significantly higher proportion. An a posteriori analysis of several CPE test-taker performances showed that the one-to-one format succeeded in eliciting on average only 14 of the 30 functions; a paired candidate format, on the other hand, was able to elicit on average 26 of the 30 functions.

Conclusion

The paired face-to-face format is not the only method to offer a valid and reliable means of assessing speaking ability and some UCLES speaking tests continue to adopt the one-to-one format for sound theoretical and practical reasons (e.g. IELTS, Young Learners). However, the studies described here (and others in progress) provide useful evidence to support the view that the paired format offers candidates the opportunity to produce a rich and varied sample of language for assessment purposes.

References and further reading

Bygate, M (1988): *Speaking*, Oxford University Press

ffrench, A (1999): 'Study of Qualitative Differences between CPE Individual and Paired Speaking Tests', Internal UCLES EFL Report

Hughes, A (1989): *Testing for Language Teachers*, Cambridge University Press

Ross, S and Berwick, R (1992): 'The discourse of accommodation in oral proficiency interviews' *Studies in Second Language Acquisition*, 14/2, 159-176

Taylor, L B (1999): 'Study of Quantitative Differences between CPE Individual and Paired Speaking Tests,' Internal UCLES EFL Report

Weir, C (1993): *Understanding and Developing Language Tests*, Prentice Hall

Young, S and Milanovic, M (1992): 'Discourse variation in oral proficiency interviews' *Studies in Second Language Acquisition*, 14/4, 403-424

Using observation checklists to validate speaking-test tasks

Nick Saville, Manager EFL Test Development and Validation Group, UCLES

This article reports on the background to developing and using observation checklists in the validation of the Speaking Tests within the Cambridge 'Main Suite' examination system. In particular, we are concerned with the application of the instruments as part of the project to revise the CPE at ALTE Level 5.

ALTE Level 1	**ALTE Level 2**	**ALTE Level 3**	**ALTE Level 4**	**ALTE Level 5**
Waystage User	**Threshold User**	**Independent User**	**Competent User**	**Good User**
Cambridge Levels				
Level 1 **KET**	**Level 2** **PET**	**Level 3** **FCE**	**Level 4** **CAE**	**Level 5** **CPE**
Basic	***Intermediate***		***Advanced***	

Figure 1: CPE within the Cambridge/ALTE Five-Level System

Milanovic and Saville (1996) provide an overview of the variables which interact in performance testing and suggest a conceptual framework for setting out different avenues of research (Figure 2). The framework has been influential in the revisions of the Speaking components of the Cambridge EFL examinations during the 1990s – including the development of KET and CAE examinations and revisions to PET, FCE and now CPE – see Saville and Hargreaves (1999) for a summary of the UCLES approach.

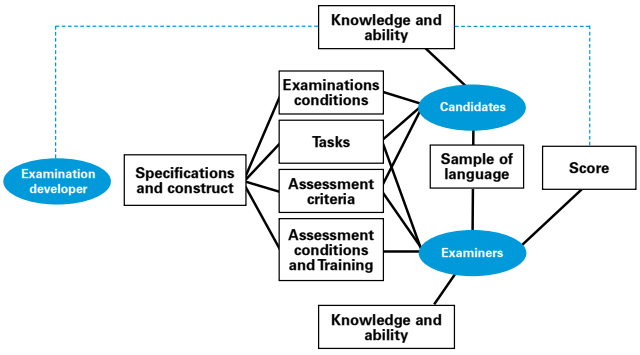

Figure 2: Milanovic & Saville (1996)

The Milanovic & Saville framework, referred to above in Figure 2, was presented at LTRC 1993 and is one of the earliest, and most comprehensive of these models (see Studies in Language Testing, Volume 3, 1996; see also Kenyon, Macnamara, etc.). This framework highlights the many factors (or facets) which must be considered when designing a performance test from which particular inferences are to be drawn; all of the factors represented in the model pose potential threats to the reliability and validity of these inferences.

The essential elements of this framework are:

- the test-taker
- the interlocutor/examiner
- the assessment criteria (scales)
- the task

For the purposes of this article it is the task which is of key importance.

The standard Cambridge approach is based on a paired format involving an interlocutor, an additional examiner and two candidates (see Lynda Taylor, p 14). The test has a number of parts and careful attention has been given to the tasks, through which the spoken language performance is elicited for each different part.

The format of the revised CPE Speaking Test can be summarised in the following specification table:

Task Format		
1	**Interviewer – Candidate**	**Interview** Verbal Qs
2	**Candidate – Candidate**	**Collaborative task** Visual stimulus Verbal instructions
3	**Interviewer – Candidate – Candidate**	**Long turns and discussion** Written stimulus Verbal Qs

Figure 3: Format of the CPE Speaking Test

While tasks are now generally recognised as of major importance in both teaching and testing contexts, in terms of test validation there is one question that has remained largely unexplored: when tasks are performed in an actual test event, how does that performance relate to the test designer's predictions or expectations based on their definition or interpretation of the construct? In validating an UCLES Speaking Test, therefore, the predicted versus actual task performance is being investigated.

UCLES EFL routinely collects audio recordings and carries out transcriptions of its Speaking Tests. These transcripts are used for a range of validation purposes, and in particular they contribute to revision projects for the Speaking Tests. For example, such transcripts were used in the FCE revision in 1996, and currently for the revision of the IELTS Speaking Test, in addition to the CPE project which is the focus of this paper.

In a series of studies within a research programme established by UCLES EFL focusing on the language of the Speaking Tests, Anne Lazaraton has applied Conversational Analysis (CA) techniques to contribute to our understanding of the language used in pair-format Speaking Tests – including the language of the candidates and the interlocutor. Her approach requires a very careful, fine-tuned transcription of the tests in order to provide the data for her analysis (see Studies in Language Testing Volume 14, forthcoming). Similar qualitative methodologies have been applied by Young and Milanovic (1992) – also to UCLES data – and by Ross and Berwick (1992) amongst others.

While there is clearly a great deal of potential to this detailed analysis of transcribed performances, there are also a number of drawbacks, the most serious of which involves the complexity of the transcription process. In practice, this means that a great deal of time and expertise is required in order to gain the kind of data that will answer the basic question concerning validity. Even where this is done, it is impractical to attempt to deal with more than a small number of tests – and therefore the generalizability of the results may be questioned. Clearly then, a methodology is required that allows the test designer to evaluate the procedures and especially the tasks in terms of the language produced by the candidates. Ideally this should be possible in 'real' time, so that the relationship of predicted outcome to actual outcome can be established using a data-set which satisfactorily reflects the typical test-taking population. The primary objective of this project, therefore, was to create an instrument built on a framework which describes the language of performance in a way that can be readily accessed by evaluators who are familiar with the tests being observed. This work is designed to be complementary to the use of transcriptions and thus to provide an additional source of validation evidence.

In the next issue of Research Notes, the work of the research team on the development and validation of the working version of the checklists will be reported. The UCLES team has been collaborating with the Testing and Evaluation Unit at Reading University, including Don Porter, Barry O'Sullivan and Cyril Weir, to produce these checklists. A paper based on this project reporting the development and early findings from the application of the checklist was presented at LTRC 2000 in Vancouver by Barry O'Sullivan and Nick Saville.

References and further reading

Bygate, M (1988): *Speaking*, Oxford University Press

Bygate, M (1999): 'Quality of language and purpose of task: patterns of learners' language on two oral communication tasks,' *Language Teaching Research*, 3/3, 185-214

Kenyon, D (1995): 'An investigation of the validity of task demands on performance-based tests of oral proficiency' in Kunnan, A J (ed) *Validation in language assessment: selected papers from the 17th Language Testing Colloquium, Long Beach, Mahwah, NJ*, Lawrence Erlbaum Associates Publishers, 19-40

Lazaraton, A (1992):' The structural organisation of a language interview: a conversational analytic perspective', *System,* 20/3, 373-386

Lazaraton, A (1996): 'A qualitative approach to monitoring examiner conduct in the Cambridge assessment of spoken English (CASE)', in Milanovic, M and Saville, N (eds.) *Performance Testing, Cognition and Assessment: selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem*, Studies in Language Testing 3, Cambridge University Press, 18 – 33

Lazaraton, A (2000) *A qualitative approach to the validation of oral language tests,* Studies in Language Testing 14, Cambridge University Press

Macnamara, T (1996): *Measuring second language performance*, Longman

Milanovic, M & Saville, N (1996): 'Introduction' in *Performance Testing, Cognition and Assessment: selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem*, Studies in Language Testing 3, Cambridge University Press, 1-17

Ross, S and Berwick, R (1992):'The discourse of accommodation in oral proficiency interviews', *Studies in Second Language Acquisition* 14/2 159 – 176

Saville, N and Hargreaves, P (1999): 'Assessing speaking in the revised FCE', *English Language Teaching Journal,* 53/1, 42-51

Skehan, P (1988): *A cognitive approach to language learning*, Oxford University Press

Upshur, J A and Turner, C (1999): 'Systematic effects in the rating of second-language speaking ability:test method and learner discourse', *Language Testing*, 16/1, 82 – 111

Weir, C (1993): *Understanding and Developing Language Tests*, Prentice Hall

Young, R and Milanovic, M (1992): Discourse variation in oral proficiency interviews, *Studies in Second Language Acquisition* 14/4, 403 – 424

Studies in Language Testing

Fairness and validation in language assessment by Antony Kunnan, volume 9 in the Studies in Language Testing Series, is likely to be of particular interest to readers of Research Notes and we reproduce below the series Editor's notes by Michael Milanovic, Deputy Director, UCLES EFL Division.

This volume is included in the Studies in Language Testing series because it represents an important statement in the on-going discussion on fairness in language testing. Fairness and its natural relationship with language test validation has been a key feature of debate in the field for the last decade. We have seen a broadening of views away from a relatively narrow focus on reliability and validity to one which recognises a complex set of relationships. Concern about this rich interaction has long been a tradition in many European language examinations. Indeed, I remember at the time of the Cambridge-TOEFL comparability study, which took place in the late eighties, John Reddaway, Secretary of the University of Cambridge Local Examinations Syndicate (UCLES) at the time, used the term 'felt fair' about Cambridge examinations in general and EFL ones in particular. Many of us did not realise how important this concept was until much later. Feeling something is fair may not be the same as it being fair but it is, perhaps, a necessary prerequisite.

Throughout the nineties, UCLES has continued the process of making its EFL examinations and tests as fair as possible. Much care and attention has gone into the materials that appear in tests. Language and topics are scrutinised, item writers and examiners carefully trained and extensive systems for monitoring quality have been enhanced. Test materials are fully pretested and examinations constructed which balance testing focus and content in accordance with published specifications. Extensive support materials are provided for candidates and training programmes for teachers. Much effort goes into developing customised test papers and procedures for candidates who are not able to deal with the conventional papers. Special circumstances, which may have disadvantaged candidates, are reported and investigated. The examination centre network is being extended continuously with about 3,000 centres where candidates can take a Cambridge EFL examination now operating throughout the world. Principles underlying performance have been investigated and instruments developed to try and understand the relationships. Much work has gone into developing and validating user-oriented scales to improve test-users' understanding of language levels and what examination scores mean in terms of performance. Many dimensions of the direct assessment of speaking and writing have been investigated and documented. Investigations into the impact of examinations have been carried out and instrumentation developed which is being shared with researchers around the world.

Given the importance of fairness and validation to the field, UCLES is pleased to add this title, edited with great care and commitment by Antony Kunnan, to the series.

Titles in the Studies in Language Testing Series are available from bookshops, or Cambridge University Press.

1 Lyle F Bachman, F Davidson, K Ryan, I-C Choi *An investigation in the comparability of two tests of English as a foreign language: The Cambridge – TOEFL comparability study,* Cambridge, 1995 (ISBN 0-521-48467-7)

2 Antony John Kunnan *Test taker characteristics and performance: A structural modelling approach*, Cambridge, 1995 (ISBN 0-521-48466-9)

3 Michael Milanovic, Nick Saville *Performance Testing, Cognition and Assessment: Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem,* Cambridge, 1995 (ISBN 0-521-484465-0)

4 Caroline M Clapham *The development of IELTS: A study of the effect of background knowledge on reading comprehension,* Cambridge, 1996 (ISBN 0-521-56708-4)

5 Alison Green *Verbal protocol analysis in language testing research: A handbook,* Cambridge, 1998 (ISBN 0-521-58635-6)

6 *Multilingual glossary of language testing terms,* Cambridge, 1998 (ISBN 0-521-65877-2)

7 Alan Davies et al. *Language testing dictionary,* Cambridge, 1999 (ISBN 0-521-658764)

8 James E Purpura *Learner strategy use and performance on language tests,* Cambridge, 1999 (ISBN 0-521-658748)

9 Antony John Kunnan *Fairness and validation in language assessment,* Cambridge, 2000 (ISBN 0-521-658748)

10 Micheline de Chalhoub-Deville *Issues in computer-adaptive testing of reading proficiency,* Cambridge, 2000 (ISBN 0-521-653800)

Forthcoming titles:

11 Catherine Elder (ed) *Experimenting with uncertainty* (ISBN 0-521-772560)

12 Cyril Weir, Yang Huizhong, Jin Yan *An empirical investigation of the componentiality of L2 reading in English for academic purposes* (ISBN 0-521-653819)

13 Kieran O'Loughlin *An investigatory study of the equivalence of direct and semi-direct speaking tests*

14 Anne Lazaraton *A qualitative approach to the validation of oral language tests* (ISBN 0-521 002672)

European Year of Languages – first call for papers

European language testing issues in a global context

The ALTE European Year of Languages conference will be held at the Universitat Pompeu Fabra, Barcelona on 5 to 7 July 2001 and organised locally by the Generalitat de Catalunya.

Main topics

Given the context of the year and the areas it focuses on, proposals under the following headings will be seen as particularly relevant:

- Assessment of younger learners and its relevance for lifelong learning
- The use of Information Technology in language testing and distance programmes
- Language testing in national education systems

Format

Proposals are requested for presentations of these types:

- **Research papers:** the presentation of completed research or papers covering theoretical topics. Paper presentations will last 30 minutes, with 15 minutes for questions and comments from the audience.
- **Symposia:** presentations involving several presenters but covering a single theme. These presentations are best suited to discussion of research and theory. Symposia will last one and a half hours and may include up to four individual presenters.
- **Poster sessions:** these provide an opportunity for the presentation of test development projects, new test development and technological innovations. Each poster presenter will be allowed five minutes during the regular program to introduce his or her project to all the participants. The poster presentations themselves will take place during a two-hour time period on one afternoon of the conference.
- **Workshops:** these provide a forum for the presentation of practical matters in a smaller group where discussion can take place and materials can be looked at and used. Such sessions would be suitable for the demonstration of technological products and materials.

Procedure for submission of proposals

Proposals should be submitted by standard mail, fax or e-mail.

Please include the following information:

1 Abstract – abstracts must be no more than one page single-spaced in length. At the top of the page please include this information:

   a the type of presentation
   b information on the presenter
   c the title of the presentation (maximum 10 words)
   d equipment required (overhead projector, projector, etc.)

Please submit four copies of the abstract, one including the name(s) of the author(s), the other three with no names on them.

2 Presenter information – please send this information about each presenter:

   a the name(s) of the presenter(s) in the order you would like them to appear in the programme, and the full name(s) of their institution(s). Underline the name of the presenter correspondence should be sent to. Please include the full name(s) of all institutions
   b title of the presentation
   c details of the person correspondence should be sent to – postal address, telephone number, fax number and e-mail address.

Address for postal submissions:

ALTE Secretariat
c/o UCLES
1 Hills Road
Cambridge
CB1 2EU

Tel: + 44 1223 553925
Fax: + 44 1223 553036
e-mail: alte@ucles.org.uk
Web site: www.alte.org

In order to be considered, proposals must be received by 31 October 2000. We will inform people of acceptance or non-acceptance of proposals by 31 January 2001.

An acknowledgement will be sent within two weeks of receipt of the proposal to the presenter who has included their details on the Standard Mail Form. If no receipt has been received by then, please contact the ALTE Secretariat.

Conference schedule

July 5 2001	Opening plenary and sessions Drinks reception
July 6 2001	Conference sessions all day Conference dinner
July 7 2001	Conference sessions all day Closing plenary

Further Information

UCLES provides extensive information on the examinations and assessment services referred to in this newsletter. For further information, visit the UCLES EFL website

www.cambridge-efl.org

or contact

EFL Information
University of Cambridge Local Examinations Syndicate
1 Hills Road
Cambridge CB1 2EU
UK

Tel: +44 1223 553822
Fax: +44 1223 553068
e-mail: harding.a@ucles.org.uk

For information on the ALTE five-level scale and the examinations which it covers, visit the ALTE website **www.alte.org**

or contact

The ALTE Secretariat
1 Hills Road
Cambridge CB1 2EU
UK

Tel: +44 1223 553925
Fax: +44 1223 553036
e-mail: alte@ucles.org.uk

If you would like further copies of this issue of Research Notes, or if you would like to be added to our mailing list (all registered UCLES centres receive copies automatically), please complete this form using block capitals and return it to EFL Information at UCLES. Please photocopy the form if you wish.

☐ Please send me ... extra copies of this issue of Research Notes.

☐ Please add me to the Research Notes mailing list.

Name .....

Job Title .....

Institution .....

Address .....

.....

.....

Country .....

Please let us have any comments about Research Notes, and what you would like to see in further issues:

IELTS MA dissertation award

To mark the tenth anniversary of IELTS, the three IELTS partners – the University of Cambridge Local Examinations Syndicate (UCLES), The British Council and IDP Education Australia: IELTS Australia – have instituted an annual award of £1,000 for the MA dissertation in English which makes the most significant contribution to the field of language testing. The first award is currently being judged and the winner will be announced in December 2000. For subsequent years, the entry procedures and timetable for the award will be as follows:

Submission and evaluation procedures

A 1000-word synopsis of the dissertation together with a reference from your supervisor should be submitted to:

Dr Lynda Taylor
EFL Division
University of Cambridge Local Examinations Syndicate
1 Hills Road
Cambridge CB1 2EU
United Kingdom.

- The IELTS Research Committee will review the submissions and short list potential award winners.
- For all short-listed dissertations a further reference will be requested by the Committee together with a full copy of the dissertation.
- The Committee's decision will be final.

Time-table

The following time-table will apply:

August	Deadline for submissions to UCLES
October	Deadline for submission of further references and copies of short-listed dissertations
December	Announcement of award