Research Notes

Contents

Editorial Notes

Welcome to issue 32 of *Research Notes*, our quarterly publication reporting on matters relating to research, test development and validation within Cambridge ESOL.

In this issue we focus on the skill of listening, the receptive skill which features in all of our language assessments. We discuss issues relevant to testing listening comprehension including establishing the nature of listening ability in a second language; the impact of technology, including the computer-based testing of listening; the writing of listening test items, including establishing sources of item difficulty and the nature of vocabulary in Listening tasks across English for Specific Purposes and General English tests.

In the opening article, Ardeshir Geranpayeh and Lynda Taylor describe the development of listening tests in Cambridge ESOL's examinations from 1913 to the present day, covering the nature of listening ability along three dimensions of a socio-cognitive framework: individual characteristics, extra-contextual factors and internal cognitive processing. They consider some of the issues with regard to assessing listening, such as the interplay of cognitive and contextual features.

The following three articles are concerned with producing and evaluating the test items used to asses candidates' listening ability. Kate Ingham describes Cambridge ESOL's training program for new item writers. All item writers undergo both general and paper-specific training; training activities for new and established item writers for the Listening component of the new International Certificate in Financial English (ICFE) are described. Next, Dittany Rose reports on a study which investigates whether vocabulary in the First Certificate in English (FCE) Listening paper is more like spoken or written language. Rose compares lexical density and word frequency patterns in this General English paper versus corpora of exam materials, source texts and native speaker material. Ardeshir Geranpayeh explores sources of difficulty for test items in a General English Listening test using Differential Item Functioning (DIF). This procedure is used to show how tests are fair to candidates and as free from construct irrelevant variables as possible. This article investigates whether age is a source of unfairness in the Certificate in Advanced English (CAE) Listening paper.

Ed Hackett then reports on how paper-based listening tests are adapted for computer-based delivery. Hackett presents some key issues in adapting paper-based tests such as displaying items and determining how candidates respond to questions, focusing on the delivery of both Business English and General English exams.

We then summarise the latest publications of interest and report on the 2007–8 ESOL Staff Seminar programme, followed by conference reports. Finally we list the 200 *Research Notes* articles available to download, the latest information on the IAEA conference Cambridge Assessment is hosting in September and the call for IELTS funded research proposals.

Editorial team for Issue 32: Fiona Barker, Ed Hackett and Kirsty Sylvester.

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

Examining Listening: developments and issues in assessing second language listening

ARDESHIR GERANPAYEH RESEARCH AND VALIDATION GROUP
LYNDA TAYLOR ESOL CONSULTANT

Introduction

Cambridge ESOL examinations have a long tradition of testing second language (L2) listening comprehension ability dating back almost a century to the introduction of the Certificate of Proficiency in English (CPE). Weir (2003:2) reports that a half-hour Dictation section formed part of the Oral paper in the first administration of CPE in 1913. This Dictation section survived the 1934, 1938, 1945, 1953, and 1966 revisions of CPE (Weir 2003:2–24). A Dictation section also formed part of the first specification of the Lower Certificate in English (LCE) when it was introduced in 1939. The inclusion of a dictation component in both tests undoubtedly reflected contemporary approaches to foreign language teaching and learning; the 'grammar-translation' method was typical at that time, and dictation skills, which combined listening, writing and knowledge of the language systems, fell within this paradigm. However, under the influence of advances in linguistics and language pedagogy during the 1960s, a dedicated Listening Comprehension Paper was introduced in LCE in 1970, replacing the earlier dictation section. The first listening comprehension test was approximately 40 minutes long. Candidates listened to an examiner reading aloud a set of passages at the front of the examination room. Passages were read aloud twice, the second time with some pausing, and candidates wrote down their answers to printed comprehension questions, including some items in multiple choice format. A similar reading aloud listening comprehension paper found its way into the revised CPE in 1975. At around the same time LCE was revised and renamed the First Certificate in English (FCE).

Defining the nature of L2 listening ability

Questions of what constitutes an 'authentic' or 'valid' approach to testing second language listening comprehension ability have long been debated and different historical periods have taken different stances depending on the prevailing approach to describing language and the nature of language proficiency. The use of dictation tests in CPE and LCE referred to above, for example, reflects one view of the nature and importance of listening ability. The inclusion in CPE from 1913 to the late 1930s of a 1.5 hour paper on English Phonetics indicates that knowledge of what words sounded like and how they were produced was at one time considered an important component of language proficiency. From the late 1960s onwards, however, English language teaching, learning and testing saw a marked shift away from a focus on *knowledge about how the language system works* towards an emphasis on the *ability to use language*. The communicative language teaching paradigm of the 1970s aimed to teach language as *a means for communication* rather than as a *system for study*, and this view was increasingly reflected in approaches to the assessment of L2 listening.

Today, we generally understand L2 listening proficiency to involve the ability to process acoustic (and possibly visual) input and to use this input to construct some sort of mental representation which in turn may be the basis for some type of spoken or written response. Taking a socio-cognitive perspective, Weir (2005:45) places acoustic/visual input as one of several core *executive processes* which the current research literature suggests are essential if we wish to develop a theoretically grounded and empirically oriented *cognitive processing* framework for L2 listening. Other executive processes, including goal-setting and monitoring, combine with *executive resources*, such as language and content knowledge. The internal mental processing dimension of listening is also shaped by a broad set of external *contextual factors* covering elements of the setting for the listening task (e.g. purpose for listening, time constraints, conditions for test delivery) and the *demands* the listening task makes on the language user in terms of linguistic variables (e.g. lexis, functions) and other variables to do with the acoustic input (e.g. speech rate, variety of accent, acquaintanceship with speaker). The *individual characteristics* of the listener – physical, psychological and experiential – will also help to shape the nature and outcomes of the listening experience.

These three dimensions – individual characteristics, external contextual factors, and internal cognitive processing – constitute three components of a socio-cognitive framework for developing and validating tests of L2 listening; they provide us with a helpful way of analysing and understanding aspects of different listening tests in terms of their context and cognitive validity. In the remainder of this article we shall consider some specific developments and issues associated with Cambridge ESOL's approach to assessing L2 listening.

The impact of technology on listening assessment

The nature and quality of acoustic input in listening tests is an aspect that has been most susceptible to changes in technology over the past 50 years. Until the 1970s, the acoustic input could only be delivered by a human speaker reading aloud a passage to a group of test takers in the examination room. But by 1984 the growing availability of tape recorders had led to the introduction of recorded listening material as part of the revision of FCE and CPE. The revised listening tests used simulated (rather than authentic) recordings of radio news, situational dialogues and

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

announcements, and at the same time incorporated charts, diagrams and picture prompts as the basis for test items. By the 1990s, advances in audio technology were providing better facilities for the administration of the listening comprehension tests. Cumbersome reel-to-reel tapes were replaced by the smaller, more convenient tape cassettes. Some years later, Compact Discs (CDs) were introduced, significantly improving the quality of listening tests.

Such technological advances have direct implications for issues of test validity and fairness in the context of listening assessment. Use of recorded listening input on cassette or CD aids standardisation of test administration, and removes the variability (and potential threat to reliability) often associated with a human reader; this is an important consideration in large-scale testing. But even if the recorded material is standardised in this way, the acoustic suitability of the room in which the listening comprehension test is taking place may impact on the performance of test takers. The nature and quality of the play-back equipment (i.e. cassette recorder, PA system, language laboratory, computer) is clearly important. It is quite reasonable to speculate that a candidate who listens to the recorded material via headphones is likely to perform differently to one who listens to the same input via loudspeakers in a large hall. More recently, the advent of wireless headphones introduces a new set of listening conditions. All these issues merit investigation to establish the potential impact of variability in aspects of the administrative setting of listening tests.

The use of technology, particularly in computer-based testing, also allows us to explore and develop new item types. This may in turn prompt us to review and expand our understanding of the listening construct, or it may enable us to test aspects of the listening construct that were not previously possible, e.g. the inclusion of more interactive/integrative tasks. In a recent review of available resources for English for Specific Purposes (ESP) testing, for example, Douglas (2007) advocates incorporating Podcasts into tests; he argues that these are becoming increasingly popular among US students as a means of social interaction and academic study on campuses, and are thus worthy of consideration.

The interplay of cognitive and contextual factors

As discussed above, the socio-cognitive perspective expressed in the framework proposed by Weir (2005) distinguishes between internal mental processes and external contextual features. In a language test, of course, there exists a close relationship between these two, as well as with how performance on the test is marked or scored (scoring validity). Weir describes this interplay in the following way:

> 'There is a symbiotic relationship between context- and theory-based[1] validity and both are influenced by, and in turn influence, the criteria used for marking which are dealt with as part of scoring validity…' (Weir 2005:20).

One of the places where matters of context and cognitive validity overlap in listening tests is in the issue of how many times the listening input is heard by test candidates. Should the recording be played only once – in an attempt at 'authenticity', i.e. replicating listening as we tend to experience it in the non-test context where it is ephemeral and we rarely get a 'second chance'? Or should it be played twice (or more) – given that the listening test context has an inherent artificiality to it, i.e. it lacks many of the visual and other support features that typically accompany the listening experience outside the test context? Interestingly, 'second chances' may be more common than we think: in interactive dialogic talk, there is usually the chance to ask an interlocutor to repeat something, while technology in the home, education and society nowadays make it increasingly easy to 'listen/watch again' (see, for example, the BBC Radio digital audiofiles and iPlayer on the BBC website).

Cambridge ESOL examinations use both once-only and twice listening formats across different tests and tasks. A convincing case can be made for both approaches, depending upon factors such as test purpose, cognitive demand, task consistency, sampling and practicality, all of which reflect the need to balance competing considerations in test design, construction and delivery. A twice listening format is used in most Cambridge ESOL examinations for a variety of reasons. One reason relates to the fact that listening to a recorded text in the context of an examination is clearly different from the listening experience that typically takes place in the world beyond the test. Relevant factors in the testing context include the absence of paralinguistic and contextual information, the time needed to normalise to speaker accents and speech patterns, and the sequential nature of the listening test task (see Boroughs 2003:336, for further discussion). Another reason concerns the need to ensure consistency across test tasks and levels (see below for an example of this with CAE). Playing the listening input twice also helps to minimise the impact of noise disturbances during live test administrations. Clear and careful guidelines are laid down to ensure minimum standards are met for facilities at test venues; in practice, however, it is not possible to guarantee that all test centres have ideal room venues or state of the art equipment for conducting listening examinations. In addition, unexpected noise may occur at any moment during the listening test (e.g. due to road/air traffic, building works, or even a candidate coughing); this can be intrusive and/or disruptive and risks impacting on candidate performance.

A test's origins or 'heritage' also understandably shape its design. The original LCE/FCE and CPE Listening tests involved the examiner reading aloud – *twice* – a set of passages on which the candidates then answered comprehension questions. The introduction in 1984 of recorded listening test material – also played *twice* – thus balanced innovation and continuity. Being given the opportunity to hear everything twice reflected, and continues to reflect, a concern for fairness in a large international market where the ability to conduct listening tests in optimum quality conditions may vary due to local constraints; in that respect it could be considered a 'virtue'. Striving to be fair to candidates is something that Cambridge has always considered a high priority.

[1] More recent versions of the socio-cognitive framework use the term 'cognitive validity' rather than 'theory-based validity'.

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

When the Certificate in Advanced English (CAE) was developed in the late 1980s, a once-only listening task was introduced in Part 2 of the Listening paper. At that time the communicative approach in teaching and testing was in full swing; ‘authenticity’ was a driving factor and this came to be reflected in various features of the original CAE design. As a brand new test, CAE sought to mirror developments in English language teaching and to have a positive impact back into the ELT community – reflecting Cambridge’s long-established concern for *consequential validity* (see Weir 2005). While maintaining a family resemblance to FCE and CPE, CAE set out to be innovative and differed quite markedly from its older brother and sister in certain respects: e.g. a mandatory paired Speaking test; new task formats for Reading and English in Use (some focusing at the discourse level for the first time); a compulsory task for Writing with textual input rather than a simple one-line prompt. The once-only listening task in CAE Part 2 (1 out of 4 tasks) was, it was believed, an acknowledgement that in much of our real-world listening we typically only hear things once; the once-only format was also well-established in the International English Language Testing System (IELTS). Any potential negative impact of hearing the text only once in slightly adverse conditions was minimised by designing the listening input with some internal repetition and by creating items focusing on explicit and easily accessible information. The parts that were not recycled were not essential to the comprehension of the text and if they were, that information was already easy to get hold of, e.g. numbers, or high-frequency, low-level words. The once-only listening task has endured since the introduction of CAE in 1991 but it has proved necessary to reassure test takers – and perhaps their teachers – that, although the text is heard once only, this does not make the task excessively demanding. The CAE Teaching Resource on the Cambridge ESOL website (www.cambridgeesol.org/resources/teacher/cae.html) states under ‘CAE Handbook, up to and including June 2008’ that ‘Part 2 is only heard once but there is plenty of time to write your answers as you listen. Key information is also rephrased and repeated within the text so you can confirm your answers as you listen.’

Despite this, however, from December 2008 Listening Part 2 of CAE will be modified to be played twice instead of just once, as part of the latest updating of FCE and CAE.[2] This will make CAE more consistent with CPE and FCE and will remove any ambiguity that may arise from listening to some tasks twice and some only once within the same test. For further discussion of the twice listening format and other issues, see Rod Boroughs’ chapter on the 2002 revision of CPE Listening in Weir and Milanovic (2003).

As mentioned earlier, the listening texts in the IELTS test are heard only once, in contrast with the majority of the ESOL tests described above. But IELTS has a different history from the other Cambridge tests and this has inevitably impacted on in its format and development. It was not originally a Cambridge ESOL test in the way that CPE/CAE/FCE are, and the once-only listening principle was established very early on its design and development (see Davies 2008). Brendan J Carroll (1978) drew up the initial specifications for ELTS (later IELTS) which had deep roots in the needs analysis and ESP movement of the 1970s. Test design therefore reflected the desire for an assessment instrument that would mirror the sorts of skills and tasks test takers would encounter in the fairly narrowly-focused Target Language Use (TLU) domain of study/training. The once-only listening reflected that same design principle, just as the pre-1989 Speaking and pre-1995 Writing modules for IELTS were directly linked to the Reading paper in an integrated way. The once-only listening variant also permits more texts and types of listening activity to be sampled within the test administration time available; this in turn allows for a larger number of test items and thus more response data to be gathered. Sampling and test length, in terms of range and number of items, are understandably constrained if all the listening input has to be repeated. Breadth of content sampling and quantity of response data are important considerations for IELTS because the test reports a modular Listening band score as well as an overall band score. In connection with this, it is also important to note that IELTS is not a level-based exam like FCE or CPE but measures across a fairly broad proficiency continuum. The 40 test items in the IELTS Listening Module are written with the aim of ranging and discriminating across a wide range of proficiency levels. For level-based exams like FCE or CPE, on the other hand, listening items are written with the specific objective of testing the candidate’s ability at a fairly well-defined and targeted proficiency level (i.e. B2 or C2 of the CEFR) and a smaller number of high-functioning test items will suffice.

Aside from the issue of once-only or twice listening, there are of course various other issues that impact on the interplay of cognitive and context validity in listening tests. These include issues such as speech rate, variety of accents, degree of acquaintanceship, number of speakers and gender of the interlocutor, all of which can have implications for the design and format of listening tests. Over recent years, for example, there has been a gradual shift in views about the inclusion of accents in listening tests. A generation ago the accents found in listening tests were predominantly British RP (received pronunciation), or Standard American English in the case of TOEFL. Nowadays, with the widespread use of English around the globe and increased exposure to local, regional and international varieties, there is greater willingness to consider including different accents in the same test. Some listening tests use a variety of native speaker accents i.e., British, Australian and North American English. Even within one national variety one may argue for the inclusion of regional native speaker varieties such as Welsh, Scottish, Cornish, Birmingham, or Liverpool accents. This debate is ongoing (see for example the exchange between Jenkins and Taylor in the English Language Teaching Journal, January 2006). The issue touches directly upon the areas of *scoring*, *consequential* and *criterion related* validity, as well as on *context* and *cognitive* validity. Inclusion of more accented varieties on context, cognitive and consequential validity grounds has to be carefully balanced against the risk of introducing test bias which is well recognised as a threat to test validity. This is where the Differential Item Functioning (DIF)/BIAS studies become so important and relevant in the

[2] See *Research Notes* issue 30.

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

continuous validation of language tests and we report separately on a DIF listening study later in this issue.

Conclusion

In this article we have tried to outline a brief history of the assessment of L2 listening in the Cambridge ESOL exams and to highlight a few of the key developments and issues relevant to testing this particular skill. The selected issues raised here have been contextualised within a socio-cognitive framework for developing and validating tests; we believe that such an approach, with its core components of test taker characteristics, cognitive validity, context validity, scoring validity, consequential validity, and criterion-related validity, provides testers with a useful and coherent framework for considering the many different features of listening tests.

In the remainder of this issue various aspects of assessing second language listening are covered starting with developing the question papers themselves – see Ingham's article on training question paper writers. Once used in a live administration, tasks and items testing listening comprehension can be analysed in various ways, both routinely and for specific research projects, as described in Rose's article comparing vocabulary in FCE Listening texts with the original radio programmes and Geranpayeh's article which explores sources of item difficulty in a CAE Listening question paper using a specific statistical methodology. The final article following this issue's theme considers how listening test items are adapted for computer-based delivery after successful paper-based delivery (see Hackett's article). The list of offprints at the end of this issue contains a number of other articles relating to testing listening which can be downloaded from our *Research Notes* website.

We have recently started work on an edited volume in the Studies in Language Testing series which will explore in much greater detail the many issues associated with assessing listening comprehension ability. The volume will be a companion to the recently published *Examining Writing* by Shaw and Weir (2007) and to two other skills-focused volumes on *Examining Reading* and *Examining Speaking* which are currently in preparation and planned for publication in 2009.

References and further reading

Boroughs, R (2003) The change process at the paper level. Paper 4, Listening, in Weir, C J and Milanovic, M (Eds), *Continuity and Innovation: Revising the Cambridge Proficiency in English examination 1913–2002*, Studies in Language Testing volume 15, Cambridge: UCLES/Cambridge University Press, 315–366.

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

Carroll, B J (1978) *Specifications for an English Language Testing Service*, London: The British Council.

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

Douglas, D (2007) *Technology and the construct of language for specific purposes*, paper presented at the 40th annual meeting of the British Association for Applied Linguistics, Edinburgh.

Jenkins, J (2006) The spread of EIL: a testing time for testers, *ELT Journal*, 60/1, 42–50.

Shaw, S D and Weir, C J (2007) *Examining Writing: research and practice in assessing second Language writing*, Studies in Language Testing volume 26, Cambridge: UCLES/Cambridge University Press.

Taylor, L B (2006) The changing landscape of English: implications for English language assessment, *ELT Journal*, 60/1, 51–60.

Weir, C J (2003) A survey of the history of the Certificate of Proficiency in English (CPE) in the twentieth century, in Weir, C J and Milanovic, M (Eds), *Continuity and Innovation: Revising the Cambridge Proficiency in English examination 1913–2002*, Studies in Language Testing volume 15, Cambridge: UCLES/Cambridge University Press, 1–56.

— (2005) *Language Testing and Validation: An evidence-based approach*, Basingstoke: Palgrave Macmillan.

Weir, C J and Milanovic, M (2003) (Eds) *Continuity and Innovation: Revising the Cambridge Proficiency in English examination 1913–2002*, Studies in Language Testing volume 15, Cambridge: UCLES/Cambridge University Press.

The Cambridge ESOL approach to Item Writer training: the case of ICFE Listening

KATE INGHAM ASSESSMENT AND OPERATIONS GROUP

Introduction

Cambridge ESOL has developed a framework for the training and development of the externals with whom it works in partnership. The framework has the acronym RITCME: Recruitment; Induction; Training; Co-ordination; Monitoring and Evaluation (see Figure 1).

In its application to writers who produce material for Cambridge ESOL question papers, the objectives of the 'RIT' part of the process are to ensure that new Item Writers have a suitable professional background, and receive training in the skills needed and information on the processes involved. In the longer term, the aim is to monitor item writing acceptance rates and to evaluate success of the item writing team on each paper. The Co-ordination, Monitoring and Evaluation stages provide the opportunity for dialogue between Cambridge ESOL and the Item Writer. This article provides an overview of the RITCME framework and describes the training of Item Writers for the International Certificate in Financial English (ICFE) Listening paper.

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

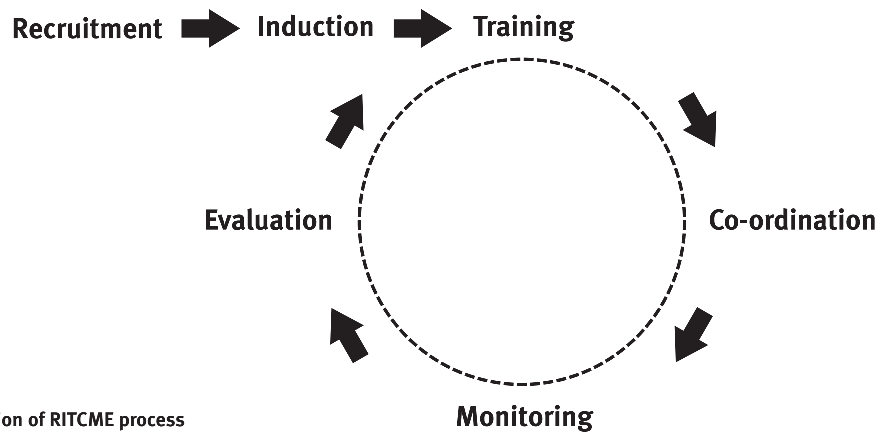


Figure 1: Illustration of RITCME process

Recruitment and Induction

The first two stages of RITCME, *Recruitment* and *Induction*, are dealt with via correspondence. Item Writers are asked to complete an application form providing details of their professional background and experience in order to ensure that they meet the Minimum Professional Requirements (MPRs) for Item Writers. MPRs exist for Cambridge ESOL's external resource personnel, such as Examiners and Presenters. The professional requirements for Item Writers include a degree and an ESOL qualification and five years' teaching experience. Some familiarity with materials production is also required, as is some involvement in preparing students for Cambridge ESOL examinations; writing and publishing experience is also desirable. In the second part of the Recruitment stage, potential Item Writers complete an introductory task on paper in order to screen them for professional suitability. Assuming that applicants are successful at this initial task, they move to the Induction stage. Induction for Item Writers requires interested applicants to read background information on the Question Paper Production process, including documentation on what is expected of Cambridge ESOL Item Writers (e.g. 'At Editing meetings, Writers are asked to contribute with suggestions for improving their own material and that of other writers on the team'). Potential writers are also sent a pre-training task relating to text selection and adaptation.

Training

Stage Three, *Training*, takes the form of a training weekend held in Cambridge which covers general issues and also ones specific to different types of papers. Cambridge ESOL Item Writer trainers work with between twelve and sixteen trainees, as detailed below.

Description of a generic training weekend

The training weekend agenda begins with an overview of Cambridge ESOL examinations and an introduction to the principles of test design and production. As part of this process, potential writers are introduced to typical test tasks (e.g. multiple-choice, productive, matching) and to the basic terminology used to describe test questions which appears in the *Item Writer Guidelines* for each paper. Each session on the techniques of writing particular item types typically lasts for two hours and includes not only input from the trainer but also group activities drawing on the ideas and experience of the participants.

To take an example, a session introducing the writing of multiple-choice questions begins by highlighting the following vocabulary in order to equip participants with the terminology for later discussion:

The answer to a multiple-choice question is referred to as

A a distractor.

B an option.

C a key.

D an item.

The answer to the above question is C. Choices A, B, C, D above are all called *options*. The question itself (*The answer to a multiple-choice question is referred to as …*) is known as the *stem*. The stem and options together are referred to as the *item*. The correct answer to the item – C in the case above – is called the *key*. The incorrect answers – A, B, D in the example above – are known as *distractors*.

Proceeding to the content of the training session, the main objective is to consider the strengths and drawbacks of multiple-choice as a test type and to identify the features of sound multiple-choice items. Participants usually agree that the task type has the advantages (inter alia) of being familiar to nearly all candidates in most parts of the world, is extremely reliable and easy to mark. As far as the qualities of good multiple-choice questions are concerned, the stem can be a question or an incomplete sentence but should have a clear focus. All options should represent a plausible response to the proposition established in the stem and should not be answerable by general knowledge. Options should be mutually exclusive and parallel with one another in wording, length and complexity (if not, they are described as not being part of a *set* – another term that is sometimes heard at Cambridge ESOL test editing meetings). Options should not cancel each other out or refer to each other (be *interdependent*) and the key should not stand out from the other options.

Examples of multiple-choice items, such as the following, are given to the participants to consider and discuss in groups:

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

The boy took the newspaper

A because he wanted to read it.

B because he wanted to dispose of it.

C because he wanted to wrap a gift in it.

D because he wanted to remove an article from it.

When presented with this particular example, participants usually comment positively on the fact that the options are comparable in structure and focus and that they are stepped in order of length. On the negative side, the options contain too much redundant information: the message to writers in this case is to include wording common to all options in the stem (in this case 'because he wanted to') in order to avoid repetition.

The techniques of item writing for other task types (word formation, sentence completion, transformation, etc.) are also discussed in other sessions, along with any relevant terminology that may arise (e.g. *base word, double key, open key, phantom, run-on*) and exemplified with appropriate examples.

Writing for particular skills papers is also covered. Most Cambridge ESOL examinations include separate papers testing Reading, Writing, Listening and Speaking. Some also have a paper which tests Use of English.[1] Participants are introduced to how writing for each of these skills has an impact on the item type and any implications for the item writer. Writing for a Listening paper, for example, includes discussion on the general desirability for options to be fairly short in order not to impose a demanding reading load on the candidate. The prospective Item Writers are given a basic overview on writing for each language skill. Other sessions during the weekend cover text selection and adaptation, drawing on a task which the participants are asked to do in advance of the training days.

At the end of the weekend, participants complete a form giving their feedback on the training weekend and details of the papers they would prefer to work on. At a later date, writers are allocated to specific item writing teams and then receive team-specific training before they start to write. Item Writers are also invited to paper-specific training and feedback events once a year. Such events generally focus on training issues which the Chair of the paper has identified or which writers have raised.

Paper-specific training for ICFE Listening

A paper-specific training event for the Listening paper of the International Certificate in Financial English (ICFE) took place in 2007. Participants included existing writers and two new to the team. The training day was led by the Chair of the Item Writing team, an Item Writer with substantial writing experience. The training was divided into three main parts: adaptation of written sources; feedback on live test performance; and selection of texts.

ICFE is an examination developed by Cambridge ESOL in collaboration with the Association of Chartered Certified Accountants (ACCA). The target candidature for the examination is finance and accounting professionals, either pre-service or in work. Item Writers for the paper mainly have a background in writing for or teaching Business English. Writers are asked to base each script for the Listening paper on an authentic source which may be written or spoken. Material for the paper thus comes from a variety of sources: sometimes from radio or television broadcasts of financial items but also from written sources. Written sources provide a wider variety of source texts: the financial press; accountancy textbooks; company documentation, etc. An issue with the use of written texts, however, is adapting the written source to spoken language. The nature of the spoken language is a key concern for Listening test papers as the candidate's listening ability is tested by answering questions about recorded spoken language. Most Cambridge ESOL Listening question papers use scripted texts which are recorded using actors as the speakers in order to ensure high-quality reproduction for assessment purposes. Although actors can be adept at making a script sound like real life, this is made more difficult if the tapescript itself does not resemble spoken language very closely (see Rose 2008 for a discussion of related issues).

Item Writers, as ESOL professionals, are aware that differences exist between spoken and written English but do not always find it easy to adapt source texts to spoken language for the purpose of listening test production. Writing for assessment is a new professional skill for many and the techniques of this specialist type of writing can sometimes seem to override other considerations, more familiar to writers. The differences between spoken and written English are well researched and rather than viewing them in opposition, it is now considered more appropriate to view genres of written and spoken language as lying on an oral/literate continuum (Tannen 1982). This continuum places some written texts such as personal letters closer to the oral end and some spoken texts such as news broadcasts or lectures closer to the literate end. Listening texts featured in ICFE cover the range of contexts that the typical candidate might encounter: financial news broadcasts, lectures, extracts from talks, interviews, consultations and professional briefings. Many of these, although spoken, fall close to the literate end of the spectrum.

The ICFE paper-specific training session began with a summary of the actual ICFE candidature in terms of age, background, first language, work sector and reason(s) for learning English. The Chair of the paper then moved on to focus discussion on the typical sources for ICFE Listening. It was agreed that written source texts were often lexically and grammatically dense but that writers frequently use such sources for reasons of topic variety and availability. There was consensus that adapting written texts to include features of spoken English in Listening tapescripts creates a more meaningful and authentic listening experience for candidates. In preparation for the training day, the Chair had produced a worksheet on which the Item Writers were asked to list typical features of informal or neutral spoken English together with examples. This was not unknown territory for participants but provided a useful refresher that spoken language uses more active verbs, more verb-based

[1] See exam handbooks on our website www.CambridgeESOL.org

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

phrases, more predicative adjectives, has a greater predominance of cleft sentences, etc. Following this, examples of written language (some from ICFE source texts) were distributed to revise and adapt for a spoken script. For example, the following sentence:

> Advances in technology have reduced the risks and costs associated with simultaneous installation of accounting software packages.

was rewritten by one group as:

> Because technology has improved, it's less risky than it used to be to install two accounting packages at the same time, and it doesn't cost so much either.

Training based on candidate performance in a live ICFE Listening paper

The training day moved on to consider performance of ICFE Listening items in the live test and at pretesting. The Chair gave an overview of performance to date. This was followed by an opportunity to consider the most recent live test paper. Participants listened to individual items and were asked to predict which ones candidates had found most challenging, and which distractors within items had worked best.

Statistical information plays an important part in evaluating the quality of items in receptive skill papers. Each paper has a target difficulty established through use of anchor items which have known measurement characteristics. In addition, items are expected to have minimum facility and discrimination values. The task-based approach of most Cambridge ESOL papers means that a range of values for items within a task is accepted as long as the task works well as a whole. The Chair identified some items to examine closely, in order to identify relevant messages for item writing.

One item involved a discussion between two accounting colleagues, Fiona and Tom, about Parcoe Metals, one of Fiona's clients. The text was as follows:

> F: Tom, I've just had the summary of performance for Parcoe Metals
>
> T: They're new clients of yours, aren't they Fiona? How are they doing?
>
> F: Well, I think Parcoe needs to be careful. I've been looking at their performance over the last two years. I suppose the most serious thing is Return on Capital Employed. Two years ago, their results were satisfactory, but there was a significant decline in profitability last year to a level well below that for the sector as a whole. Fortunately, there's no sign of nervousness amongst the investors. Something to be grateful for.

The accompanying item (Question 7) was:

> Fiona is worried because Parcoe's Return on Capital Employed
> A has fallen for the second year running.
> B is lower than the industry average.
> C will have a negative effect on shareholders.

The difficulty of the item was 68.65, within the target difficulty range for the paper. The participants were also given classical item statistics relating to each question. The ones for question 7 are shown in Table 1.

Discussion took place as to how well the item had performed. The key (option B) was clear and had discriminated well with 95% of the 'high' group of candidates (that is, the most proficient) selecting it. The Chair clarified the meaning of the available statistical indices for the benefit of the new item writers. The *Prop Correct* (proportion correct) also known as the facility, shows the percentage of candidates answering the item correctly; in the item above this was 70%. Distractor A performed best in that 23% of the candidates found it attractive, whereas Distractor C had relatively poor 'pulling power', appealing to only 7% of the candidature.

The Chair also briefly summarised Item Discrimination, particularly for the new participants. Cambridge ESOL has minimum and maximum Item Discrimination targets for its tests. In general, the higher the item discrimination, the better an item has performed, although a range is needed within a particular paper and some statistics can only be interpreted within the context of a given test population. Measures of Item Discrimination show how successfully an item distinguishes between higher and lower ability candidates. The *Point-Biserial Correlation* shows the relationship between the candidates' performance on a single item and their performance on all items in the test, i.e. do those people who answer the item correctly also score highly on the rest of the paper? The Point-Biserial for the item above was .61. The *Discrimination Index* also reports how well items perform: the highest and lowest scoring groups of candidates are compared. The proportion of the lowest scoring candidates answering correctly is subtracted from the proportion of the highest scoring group. In the ICFE item below, the Discrimination Index is .63 (.95 minus .32). Although the figures are comparable, the Point-Biserial is usually considered preferable to the Discrimination Index as it takes into account all candidates, not only the top and bottom thirds.

The Item Writers agreed that Distractor C in Question 7 had been fairly easy to rule out due to the match between tapescript and option of 'shareholder' and 'investor' but

Table 1 : Classical item statistics for Question 7

		Item Statistics			Alternative Statistics					
Seq. No.	Scale Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Endorsing			Point Biser.	Key
						Total	Low	High		
7	1–7	.70	.63	.61	A	.23	.53	.05	–.52	
					B	.70	.32	.95	.61	*
					C	.07	.16	.00	–.25	

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

that Distractor A had required greater processing and so had acted as a greater distraction to some candidates. The implications of this for item writing were discussed. Other items were considered with this concern in mind.

Training in text selection

The third part of the day involved consideration of source texts and approaches to item writing. The Chair had pre-selected a range of texts from different sources. Writers were asked to work in groups discussing each text and its potential as a basis for an ICFE Listening task, and if so, for which task on the paper it could best be exploited. The new Item Writers were put into groups with the more experienced writers. The discussion at the end provided a degree of consensus on the best texts to use: this was down to various factors such as topic, content and style of writing. It also led to a fruitful discussion on item writing technique. The experienced writers agreed that they first skimmed a text with a view to assessing whether it would yield the requisite number of items for a particular task, whereas the new writers had thought that this would be the secondary consideration. One experienced writer demonstrated how she swiftly evaluated fairly lengthy texts by underlining potential keys with a highlighter pen.

The training day finished with the Chair providing advice on the forthcoming writing commission and offering support, particularly to the new writers, should they need to contact her with any queries.

Co-ordination, Monitoring and Evaluation

The *Co-ordination* stage of RITCME includes encouragement to writers through comment on their material and support in Editing meetings. In addition, Cambridge ESOL communicates with all of its Item Writers on an annual basis to provide overall feedback on their work and on the performance of the papers on which they write. The objective is to maintain a high standard of writing quality across the team for each paper.

Item Writing teams are monitored in respect of the amount of material commissioned versus the amount of material submitted and accepted for Editing. It is expected that 100% of material commissioned should be submitted and there is a target of at least 80% to be accepted for Editing by the team as a whole.

Paper performance is also monitored through use of statistics showing how much material has moved successfully through the pretesting stage and reached the Test Construction bank of LIBS (the Local Item Banking System; see Marshall 2006). The Chair of each paper, using these statistics together with experience from Editing meetings, writes a summary of the overall item writing team performance and an evaluation of each individual team member's contribution. This information is included on a feedback form sent to each writer for each paper on which they write. Writers are asked to read and complete the forms. There is also space on the form for each writer to request specific training or identify particular writing issues: the aim is to provide the opportunity for Item Writers to communicate any concerns directly with the organisation. The forms are returned to Cambridge ESOL and information is collated and sent to the Chair as the Monitoring and Evaluation parts of their RITCME.

Conclusion

Given the vital contribution made by Item Writers to Cambridge ESOL examination papers, it is important that a training and development structure exists. RITCME for Item Writers also provides a channel for transparent two-way dialogue between Cambridge ESOL and the Item Writer which is continuously evolving. The whole RITCME framework is discussed and reviewed on a regular basis to ensure that it meets the needs of those who use it and that it sustains consistent standards in the effective production of high-quality test papers.

References and further reading

Green, T and Jay, D (2005) Quality Assurance and Quality Control: Reviewing and pretesting examination material at Cambridge ESOL, *Research Notes* 21, 5–7.

Marshall, H (2006) The Cambridge ESOL Item Banking System, *Research Notes* 23, 3–5.

Rose, D (2008) Vocabulary use in the FCE Listening test, *Research Notes* 32, 9–16.

Tannen, D (1982) The Oral/Literate Continuum in Discourse, in Tannen, D (Ed.) *Spoken and Written Language: Exploring Orality and Literacy*, Norwood, New Jersey: Ablex Publishing Corporation, 1–16.

Vocabulary use in the FCE Listening test

DITTANY ROSE RESEARCH AND VALIDATION GROUP

Introduction

This article is based on corpus-informed research carried out for a Masters level dissertation at Anglia Ruskin University (UK). The study investigated vocabulary use in the First Certificate in English (FCE) Listening paper, to see if the vocabulary used in the listening texts was more like spoken or written language. The research questions which this study sought to answer were:

1 Do FCE listening texts, which have been based on real-world spoken texts, have different lexical densities from the real-world spoken texts?

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

2 Do FCE listening texts, which have been based on real-world spoken texts, have different patterns of word frequency from the real-world spoken texts?

Two corpora were created, one containing texts taken from FCE Listening exams and one containing the real-world texts on which the test materials were based. The vocabulary used in these two corpora were compared with each other and also with a larger corpus, the British National Corpus (BNC). The measures used for comparison were *lexical density* and *lexical frequency* as these have been shown in other studies to be good indicators of spoken-ness or written-ness. It was predicted that the process of editing texts would change the lexical density of the texts and change the frequencies of the words used. The research intended to provide insights which could inform item writer training.

The use of real-world texts in listening exams and text books

The use of texts taken from real-world contexts is a major feature of Cambridge ESOL exams, including the FCE Listening paper. There is some concern in the language testing literature, however, most noticeably from Buck (2001) about the way that listening texts, by their very nature, can suffer from the constraints of professional production. If worked on and edited as written texts, Buck suggests they are likely to become, to some extent, more like written language than spoken language. In another study, Gilmore (2004) looked into the discourse features of listening dialogues of service encounters in ELT coursebooks, comparing the textbook dialogues to real world service encounters. He looked at a number of discourse features including word length, turn taking, hesitation devices and lexical density. His findings indicated that the textbook dialogues differed 'considerably, across a range of discourse features' (Gilmore 2004).

The aim of the research described here was to see if there was any evidence that FCE Listening texts are in any way more like written language than spoken language. The research focused on linguistic features of the texts, as an understanding of these could have implications for item writer training. In particular the focus was on lexis as it is one of the key differentiators between spoken and written text and is an area covered by both Buck (2001) and Gilmore (2004). McCarthy (1998) also pointed out the 'rather scant amount of research into the kinds of vocabulary patterns that occur in everyday spoken language', as opposed to grammatical patterns which he feels have been well documented.

The First Certificate Listening test

The FCE exam is at B2 level on the CEFR and is part of the Cambridge ESOL Main Suite of examinations, a series of five exams at different levels. FCE has four skills based components – Reading, Writing, Listening and Speaking – plus one Use of English paper.

Each FCE Listening test contains the following number of tasks and texts:

Part 1	8 tasks	one text (80–110 words) and one question per task
Part 2	1 task	one text (600–675 words) and ten questions
Part 3	1 task	five texts (80–110 words each) and five questions
Part 4	1 task	one text (600–675 words) and seven questions.

Cambridge ESOL has a set of standard procedures for producing exams, which takes a minimum of eighteen months from material being commissioned to its use in a live exam paper. This is described in more detail in Marshall (2006).

The texts and tasks for the FCE Listening paper are submitted in the form of written tasks and accompanying written tape scripts. They are worked on as written artefacts and are first heard in full when they are recorded in a studio with professional actors.

When submitting material for Main Suite Listening papers, item writers are asked to base each text on an authentic context which may be written or spoken, for example, a newspaper article or a radio discussion programme. Item writers are directed to make sure that, wherever the original text comes from, the tape script is written using oral rather than written language.

Testing Listening

Real-world speech contains many ambiguities, pauses and false starts. If any of these affect comprehension the listener can clarify by asking questions. In an exam situation this is not possible and so texts and answer keys for listening tests need to be unambiguous. Added to this there are constraints of time and length and the need to mediate culturally specific references. For all these reasons real-world texts will need a degree of adaptation to make them suitable for use in a Listening test. The effects of this adaptation on the listener have been little studied and research in the area of listening comprehension has been limited, as Joan Rubin (1994) notes. One study she quotes (Voss 1994) appeared to show that pause phenomena distracted from students' comprehension of a text. However, Chaudron (quoted in Rubin 1994) and Chiang and Dunkel (quoted in Rubin 1994) have both shown that redundancy helps comprehension in high-intermediate students[1].

These findings are, however, inconclusive, as most of the studies quoted were done using invented, isolated text (Rubin 1994).

Key differences between written and spoken language

Spoken and written language have evolved different grammars and vocabularies. Summarising these, Chafe (1982) noted that spoken language is characterised by fragmentation and written language, by integration. Chafe

[1] Although it hinders comprehension in low intermediate students.

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

found that this integration was achieved by the use of methods such as nominalisation and the use of participles and attributive adjectives. In other words, it tends to be grammatical words that are lost in this process of integration, to be replaced by a greater use of information-giving words.

Chafe (1982) also realised that the way speakers interact with their audiences leads to spoken language being characterised by involvement, whereas written language is characterised by detachment. He notes the ways in which speakers create involvement: by the use of the first person, the vocalisation of mental processes, monitoring of information using discourse markers and the use of 'fuzzy' language – or language that is purposefully vague. Vague language tends to use high frequency lexis and grammatical words, such as: *things like that, or anything, and so on, loads of*.

It is important to note, however, that the differences noted above are not dichotomous. Tannen (1982) talks about the notion of an oral/literate continuum, with some written texts, such as personal letters relatively closer to the oral end and some spoken texts, such as lectures, closer to the literate end.

This article will now consider two aspects of lexis that researchers have used to define genres and to place texts at some point along this continuum: lexical density and word frequency.

Lexical Density

Halliday (1989) noted that 'written language displays a much higher ratio of lexical items to total running words' than spoken language, and went as far as to say that the defining characteristic of written language is lexical density. In its simplest form, lexical density is the percentage of lexical items (L) to total number of items (N) in a text: 100 × L/N[2].

Stubbs (1996) who considers lexical density to be 'a robust method of distinguishing genres', carried out a study in which he found lexical densities for spoken texts of 34% to 58%. For written texts the figures were 40% to 65%. He concluded that the 'clearest difference is not between written and spoken language but between spoken genres'. He also noted that genres where no feedback was possible (that is, monologues such as radio commentary) had higher levels of lexical density – from 46% to 64% – and those where feedback was possible (that is dialogues such as radio discussions) had lower levels – from 34% to 44%.

It is hypothesised that the genre of radio programmes comes more towards the literate end of the oral/literate continuum than some other modes of speech such as language in action conversations. This is because radio interviews are often partially planned. Whilst the interviews may not be scripted, the interviewee may have been briefed about the topics likely to be covered. This means that they will be less likely to need thinking time as they are talking and will be less likely to make errors or need to restate a point than in, say, a conversation with friends. The participants will also be aware that they are broadcasting to people who have varying degrees of background knowledge about the topic. This means that the participants are likely to clarify and explain more than if they knew they were talking to, for example, a colleague in the same field.

It is probable, therefore, that on the oral/literate continuum this genre will be somewhat closer to the literate end than other oral genres.

Word frequency and corpora

In order to compare individual texts to the language as a whole, a large amount of data from corpora is needed. This study uses data from the British National Corpus (BNC) which is a corpus of British English set up by a collaboration of Oxford University Press, Lancaster University and other organisations.[3] There are 100 million words in the BNC, 10% of which are spoken. The spoken texts are of two main types: conversational (40%) and task oriented (60%).

Scott (1996) states that; 'Word frequency information is very useful in identifying characteristics of a text or of a genre'. The top of any frequency wordlist of suitable size usually contains a small number of high frequency words – these will mostly be function words. Slightly further down the list the first high frequency content words will appear: 'Typically among the first content words in a wordlist are verbs like know, said, think' (Scott 1996). Nouns tend to come much further down the list. The bottom of any list is likely to contain a large number of 'hapax legomena'; words which only occur once.

Interesting patterns emerge when separate wordlists are run for written texts and spoken texts. McCarthy (1998) for example, reports on wordlists from 100,000 words of written data and the same amount of spoken data. All the top 50 words in the list for the written texts are function words, which may seem surprising considering what we have said earlier about the heavy lexical load of written language. But it is precisely this phenomenon that gives written language its density. It makes use of a far greater range of lexical items of lower frequency and therefore has a greater lexical load.

The spoken list, on the other hand, appears to have a number of lexical words in the top fifty; *know, well, get/got, go, think, right* (McCarthy 1998). McCarthy indicates that most of these high frequency words are not used lexically in all cases, often being part of discourse markers such as *you know* or *I think* (see also McCarthy and Carter 2003).

Methodology

This research took place in several stages. First, suitable Listening tests were selected. Then, lexical densities were calculated and word frequency lists created.

Selecting Listening texts

Texts from past FCE papers which were sat between June 2002 and December 2005 were used for this research. The two parts of the FCE Listening paper with longer texts (Parts 2 and 4) were considered and nine were selected from those available for investigation.

[2] There are many suggested ways to calculate lexical density (see, for example, O'Loughlin 1995, Stubbs 1986, Ure 1971).

[3] See www.natcorp.ox.ac.uk

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

Table 1: Original and adapted texts and their source programmes

Radio text name	No. of Speakers	Programme	Exam text name	No. of Speakers
Parakeets	4	Natural History Programme	Birds	1
Bones	3	Natural History Programme	Dinosaur Discovery	2
Costa Rica	2	Natural History Programme	Cable Car	2
Urban Wildlife	4	Natural History Programme	Nature Reserve	1
Lara Hart	2	Books and Company	Lucy Bray	2
Patricia Routledge	2	Desert Island Discs	The Actress	2
Victoria Beckham	2	Womans Hour	Cool Pepper Band	2
Janet Ellis	1	The Musical Side of the Family	Celebrity Families	2
James Dyson	2	Desert Island Discs	David Dickinson	2

Two corpora were formed, one containing nine exam texts and the other containing the nine corresponding radio texts on which the exam texts were based, as shown in Table 1. The exam texts consist of the tape scripts, each of 600–700 words. The radio texts are transcriptions from recordings of radio programmes, from 600–5000 words each. In total there are 6,149 words in the exam texts corpus and 21,277 in the radio texts corpus. Table 1 lists the names and number of speakers of the original radio texts and their adapted exam texts together with the source radio programmes.

The eighteen texts were imported into two composite text files, one for the original radio texts and one for the exam texts. Both text files were then analysed as described below.

Calculating lexical density

The following categorisations were used:

Grammatical items

- All proforms (*she, it, someone*)
- All determiners (*the, some, any*)
- All prepositions and conjunctions
- All examples of the verbs '*to be*' and '*to have*'
- Numbers
- Reactive tokens (*yes, mm*)
- Interjections (*gosh, oh*)
- Lexical filled pauses (*well, so*)
- Non-lexical filled pauses (*er, erm*)
- Discourse markers (*you know, I mean, like*) counted as single items.

Lexical items

- All nouns, main verbs, most adjectives and adverbs counted as lexical items.
- Verbs '*do*' and '*go*' counted as lexical only where used as a main verb.

A manual approach to counting was used after Zora and Johns Lewis (1989 quoted in O'Loughlin 1995). In order to eliminate issues of variable text length, two hundred words were analysed, two hundred words into the text – that is words numbered from 200 to 400 in each text.

In this article one pair of texts is exemplified: Costa Rica/Cable Car. Costa Rica is a subsection from a longer radio magazine programme on the topic of nature. It is an interview with a man who set up a cable car in a jungle. The adapted exam text Cable Car retains the interview format. See Figures 1 and 2 for examples of the manual assigning of lexical items to these texts. (Items in bold are lexical items).

Figure 1: Extract from Costa Rica

> **Right now** there's a **big problem** with **deforestation** in **Costa Rica** and one of the **things** that we **need** to do is to **provide education** and we have a **great opportunity** here. We've **got** an **education programme** in **place** where we will **bring students** in, **free** of **charge** and **tell** them about er the **canopy** and why it should be **saved**…

Figure 2: Extract from Cable Car

> … need to do to **stop** that is to **provide education**. We've **got** a **programme** in **place** where we will **bring students** in from all over the **world** and **tell** them about the **forest** and they can **see** for **themselves** why it should be **saved**.

Creating word frequency lists

Wordlists were made by running the two corpora through WordSmith Tools (Stubbs 1996). The resulting exam texts wordlist and radio texts wordlist were compared to published British National Corpus (BNC) lists for spoken and written language (Leech, Rayson, Wilson 2001).

The wordlists were then used to make *key word* lists in WordSmith Tools, using the larger composite radio text as a reference text. The KeyWord tool finds words which are significantly more frequent in one text than another. If a word is unusually *in*frequent in the smaller corpus (the exam texts here), it is said to be a negative key word and will appear at the end of the list.

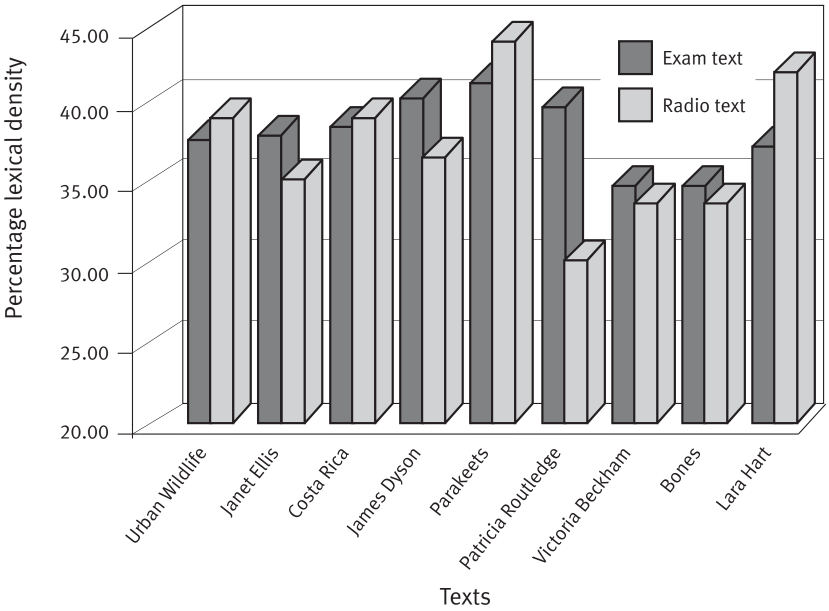
Concordances were then run on selected words, so that their usage in both sets of texts could be studied in more detail.

Results: lexical density

As can be seen from Figure 3, all texts ranged from 30% to 44% lexical density. This is lower at the bottom and top of the range than Stubbs' (1996) finding for spoken texts (34% to 58%). At the upper end, this difference can be accounted for, as most of the texts studied here are dialogues and would not be expected to have particularly high lexical densities. The presence of results which are lower than 34% could, however, suggest that the method for calculating lexical density used in this study created different results from Stubbs' method.

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

Figure 3: Lexical density of radio and exam texts



None of the texts analysed, whether exam or radio texts, have a lexical density greater than 44%, even though some of them are monologues where there is no feedback. This would suggest that all these radio texts are dialogic in some way, with speakers regarding the listeners as involved in the interaction to some extent, even though there is no option for actual feedback.

It is hard to see a particular pattern when comparing the exam texts to the radio texts; some exam texts have higher lexical density than the corresponding radio texts (five texts) and some radio texts have higher lexical density than exam texts (four texts). Overall though, the exam texts have a slightly higher lexical density. The average is 37.5% as opposed to 36.8% for the radio texts.

An independent t-test for significance was carried out using SPSS©. There was no significant difference found between the conditions (t=.443, df= 16, p=.663, two tailed). This shows that the difference between the mean lexical densities of the exam texts and the radio texts is not significant to 95% probability. That is to say, it is reasonable to assume that the differences in mean are attributable to chance.

What is noticeable, however, is the range of densities in the texts. The difference between the highest and lowest densities on the radio texts is 13.5%. On the exam texts this difference is only 6.5%, so it seems there is a tendency for the radio texts to have more variation in lexical density and the exam texts to conform to an average density.

Results: word frequency

Table 2 shows the top 50 words in radio texts, exam texts and, for comparison, the BNC spoken corpus and BNC written corpus.

If we take a closer look at the top ten items, we can see that all of the top ten words in all four corpora are function words. The same top ten words appear in the radio texts as in the BNC spoken corpus, although in a different order. These results indicate that the corpus of radio texts may be as representative of spoken English as the BNC, and

Table 2: Top 50 words in radio texts, exam texts, BNC spoken corpus and BNC written corpus

	Radio texts	Exam texts	BNC spoken	BNC written
1	the	the	the	the
2	and	and	I	of
3	a	a	you	and
4	I	I	and	a
5	you	to	it	in
6	to	of	a	to
7	it	in	’s	is
8	of	it	to	to
9	that	that	of	was
10	’s	you	that	it
11	was	was	n’t	for
12	in	’t	in	that
13	’t	we	we	with
14	but	but	is	he
15	we	so	do	be
16	is	what	they	on
17	on	is	er	I
18	they	’s	was	by
19	for	my	yeah	’s
20	she	for	have	at
21	have	they	what	you
22	so	about	he	are
23	very	on	that	had
24	erm	as	to	his
25	at	have	but	not
26	know	at	for	this
27	well	when	erm	have
28	with	all	be	but
29	what	me	on	from
30	think	do	this	which
31	as	like	know	she
32	do	people	well	they
33	this	be	so	or
34	there	from	oh	an
35	all	really	got	were
36	he	this	’ve	as
37	about	well	not	we
38	er	know	are	their
39	be	there	if	been
40	not	an	with	has
41	had	one	no	that
42	’ve	with	’re	will
43	really	had	she	would
44	her	if	at	her
45	when	don’	there	there
46	my	think	think	n’t
47	because	did	yes	all
48	yes	very	just	can
49	been	are	all	if
50	if	can	can	who

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

suggests that although we may consider radio programmes to be a specialised genre, they are not too narrowly defined or restricted in language use.

The top four items in the exam texts list are the same and in the same order as the radio texts list; overall nine of the top ten words are the same in both lists. The exceptions are *in*, which is at position 7 in the exams texts list and 12 in the radio list, and ‘*s* which is at position 10 in the radio list and 18 in the exam texts list.

Lexical items

There are two lexical words in the top fifty of the BNC spoken corpus *know* and *think*, whereas there are no lexical words in the top fifty of the BNC written corpus. There are four lexical words in the top fifty in the radio texts, *very, know, think* and *really*. The fact that there are more lexical words here than in the BNC top fifty can be accounted for by the smaller corpus size. Two of these, *think* and *know*, are words which are used within discourse markers: *I think*, *you know*. They are also used to vocalise mental processes, which was another feature of spoken language that Chafe (1982) noted. *Really* and *very* are words which have some overlap in meaning so it is interesting that they both appear high up on the radio texts wordlist.

All four of the lexical words in the top fifty in the radio texts also occur in the top fifty in the exam texts although the order is a little different. There are two other items which occur in the top fifty exam texts but not in the top fifty radio texts: *people* and *like*.

Filled pauses, interjections and discourse markers

These do not appear in the BNC written corpus as they are a purely spoken phenomenon. Accordingly, in the BNC spoken corpus: *er, yeah, erm, well, so, oh, no* and *yes* appeared. In the radio texts corpus *yes, well, really, so, erm* and *er* occurred. It is not possible to say from the list alone whether *so, well* and *really* are used as discourse markers or what part of speech they are, which could be investigated with concordances.

It is interesting to note the absence of *oh* from the radio texts top fifty. Leech et al. (2001) find that

> ‘Most interjections (e.g. *oh, ah, hello*) are much more characteristic of everyday conversation than of more formal/public “task oriented” speech. However, the voiced hesitation fillers *er* and *erm* and the discourse markers *mhm* and *um* prove to be more characteristic of formal/public speech. We recognise *er, erm* and *um* as common thought pauses in careful public speech. *Mhm* is likely to be a type of feedback in formal dialogues both indicating understanding and inviting continuation. In conversation, people use *yeah* and *yes* much more, and overwhelmingly prefer the informal pronunciation *yeah* to *yes*. In formal speech, on the other hand, *yes* is slightly preferred to *yeah*.’

The absence of *oh* and the presence of *er* and *erm* in the radio texts suggest that they lie more in the area of formal or public speech than conversation. This is also backed up by the much greater use of *yes* than *yeah* in the radio texts corpus. In the exam texts corpus only *so* and *well* occurred, both of which also have uses other than as interjections or discourse markers. There are no non-lexical filled pauses in the exam texts top fifty list. This shows that the exam texts are missing this element of natural speech.

Table 3: Selected key words displaying positive keyness

N	Word	Freq - Exam texts	List % - Exam texts	Freq - Radio texts	List % - Radio texts	Keyness
2	in	131	2.07	295	1.34	16.5
3	reserve	4	0.06	0	12.0	12.0
6	my	44	0.69	82	0.37	10.4
7	although	7	0.11	3	0.01	10.3
8	bones	7	0.11	3	0.01	10.3
10	especially	5	0.08	1		10.1
11	rainforest	10	0.16	8	0.04	9.3
15	wondered	3	0.05	0		9.0
17	job	6	0.09	3	0.01	8.1
18	forest	9	0.14	8	0.04	7.5
19	nature	4	0.06	1		7.5
20	some	17	0.27	24	0.11	7.5
21	college	7	0.11	5	0.02	7.2
22	birds	14	0.22	19	0.09	6.6
23	survive	2	0.03	0		6.0
26	survey	2	0.03	0		6.0
27	proved	2	0.03	0		6.0
29	ordinary	2	0.03	0		6.0
31	cities	2	0.03	0		6.0
32	cake	2	0.03	0		6.0
34	cages	2	0.03	0		6.0
36	discuss	2	0.03	0		6.0
37	directors	2	0.03	0		6.0
38	insects	2	0.03	0		6.0
39	homework	2	0.03	0		6.0
40	employees	2	0.03	0		6.0

These results suggest that the radio texts corpus is to some extent composed of more formal speech than the BNC spoken corpus. There are indications that radio interviews are, as suspected, somewhere towards the literate end of the oral/literate continuum. However, they are still representative of spoken language and do not show similarities with written language. The exam texts seem to mirror the radio texts fairly well, although there is a noticeable absence of non-lexical filled pauses.

Key words and concordances

Tables 3 and 4 show the key words displaying positive keyness (at the top of the key words list) and negative keyness (at the bottom of the key words list).

The top of the KeyWords list (Table 3) contains a number of names, for example *Maddy*, which have not been listed here. In order to avoid interference from world knowledge, Item Writers amend famous names. The replacement names will automatically come up as key words as they appear a number of times in the exam texts but not at all in the radio texts.

Most of the other positive key words in Table 3 are nouns or adjectives; *reserve, bones, rainforest, job, forest, nature, college, birds, British, city, model, research, famous, project, book, food, band, successful, cable, novel, hundred, area, royal, young*. These relate to topic, and give the text its ‘aboutness’ as Scott (1996) describes it. Looking at these words we can get a good idea of what topics are covered in the exam texts. These items do not necessarily indicate use of different words in the two texts and their appearance on the list may be a result of text length. An adapted, that is shortened, text will need to retain its core ideas and these will include key topic words. Another reason that nouns and adjectives appear as key words is when an item writer makes the decision to replace a low frequency word with a higher frequency one. For example,

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

Table 4: Key words displaying negative keyness

N	Word	Freq - Exam texts	List % - Exam texts	Freq - Radio texts	List % - Radio texts	Keyness
400	else	1	0.02	14	0.06	2.7
401	think	23	0.36	117	0.53	3.0
402	because	14	0.22	80	0.36	3.3
403	‘ll	2	0.03	22	0.10	3.3
404	then	8	0.13	53	0.24	3.4
405	anything	2	0.03	23	0.10	3.7
406	course	3	0.05	29	0.13	3.7
407	must	1	0.02	18	0.08	4.3
408	him	1	0.02	19	0.09	4.7
409	sort	3	0.05	32	0.15	4.7
410	yes	12	0.19	79	0.36	4.9
411	her	13	0.21	84	0.38	5.0
412	re	11	0.17	75	0.34	5.1
413	time	8	0.13	61	0.28	5.3
414	his	2	0.03	28	0.13	5.4
415	mean	9	0.14	69	0.31	6.1
416	very	22	0.35	132	0.60	6.4
417	er	13	0.21	93	0.42	7.1
418	you	118	1086	570	2.58	11.5
419	she	18	0.28	139	0.63	12.4
420	‘s	47	0.74	327	1.48	23.6
421	erm	6	0.09	122	0.55	31.3

in the text ‘Birds’, the word *bird* was used instead of the less familiar and much lower-frequency *parakeets* which was used in the original radio source text.

Entries at the bottom of the list (the negative keywords in Table 4) do not include names; there are few nouns or verbs and for this reason they do not give a sense of the ‘aboutness’ of a text. They do, however, give a sense of the ‘spoken-ness’ of the radio texts: *sort, yes, her, ’re, time, mean, very, er, you, she, ‘s, erm*. It is this sense which is absent from the exam texts. The negative key words are very interesting, as they are mostly of a grammatical nature. There are lexical words at the bottom of the list: *very, time, sort*, but these are either part of a commonly used discourse marker or are lexicogrammatical.

Words which showed negative keyness or high positive keyness were studied in more depth: statistics on their occurrence were tabulated and compared and concordances were run. Two entries are exemplified below in Tables 5 (the conjunction *although*) and 6 (the verb *mean*) although for reasons of space the concordances are not shown.

Positive keyness

The conjunction *although* is used more often in the exam texts than in the radio texts. It is used 7 times in the exam texts and only 3 times in the radio texts (see Table 5). Data from the BNC indicates that it is more common in written language than spoken. This may indicate that item writers or editors are adding in words which are more from a written media. However, the instances are low so it is difficult to really make any judgements based on these.

Negative keyness

The verb *mean* is used less often in the exam texts than in the radio texts (see Table 6). Six out of the nine instances in the exam texts and sixty out of the sixty nine in the radio texts are as part of the discourse marker ‘*I mean*’, which can be used to correct information or to start or continue a sentence (also see Erman 1987). It is interesting to note that in the radio transcripts this is used both in programmes with an older speaker (*Patricia Routledge*) and those with a younger speaker (*Victoria Beckham*). In the exam texts however, five of the nine examples are in the *Cool Pepper Band* text. There are no examples of this discourse marker in the *Actress* text.

Table 5: ALTHOUGH (conjunction)

BNC Speaking frequency	160
BNC Writing frequency	468
Exam text frequency	7
Radio text frequency	3
Keyness	10.3

Table 6: MEAN (verb)

BNC Speaking frequency	2250
BNC Writing frequency	198
Exam text frequency	9
Radio text frequency	69
Negative Keyness	6.1

It may be that when rewriting the text the item writer had not noticed the use of this phrase and would not have associated its use with an older speaker. Alternatively, writers may be removing these because the word count is limited.

Conclusion

If this study were to be repeated, the length of each text should be examined more closely and a methodology developed to make sure that the radio texts are all longer than the exam texts but of a similar length to each other. In this study the results of the WordList word frequency list and KeyWords analysis have been affected to some extent by this unequal text length.

The difference in Lexical Density was not found to be statistically significant between the two corpora. There was, however, less variety between all the exam texts than between all the radio texts. Or to put it another way, there is a tendency to uniformity between the lexical densities of Part 2 and 4 FCE Listening texts. This is in a sense to be expected, and to be welcomed, as it indicates that different candidates, doing different versions of the tests, get texts with similar properties. This suggests that existing item writer training and question paper production procedures help to achieve fairness for all candidates (see Ingham 2008).

On the other hand, it could be argued that one of the skills that a learner should be tested on is their ability to cope with different types of text, with different degrees of lexical density and different forms of redundancy. Further studies could usefully be carried out to look at the lexical densities of Part 1 and Part 3 FCE Listening texts to see if there is more variety over the whole test when these parts are taken into account.

When looking at word frequency there were some differences between the exam texts and the radio texts. The exam texts showed less use of filled pauses and discourse markers than the radio texts. They may also make less use

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

of vague language, which is characteristic of spoken language. There is no conclusive evidence regarding use of other categories of lexis, but overall it was noticeable how the negative keywords in the WordSmith Tools analysis felt 'spoken'. That is to say, that what had been removed in the translation of a text from a radio broadcast into an exam text, were features of a more spoken nature.

This difference may of course be entirely justified, even desirable, in the context of language assessment. There is some evidence that students' comprehension of a text may be hindered by use of pause phenomena and discourse markers. There is also a growing understanding of the concept of authenticity and the fact that adaptation for level – and other reasons such as cultural appropriacy – does not automatically 'disauthenticate' a text (see Murray 2007). Changes to listening texts, so that they are suitable for use on the FCE Listening test, are made under expert judgement and backed up by statistical evidence of the performance of the task when it is pre-tested before being used in a live administration. The tasks covered in the article all performed well in live tests with large candidate numbers and this is the clearest evidence that the Listening texts investigated here are pitched at the right level in terms of their content.

References and further reading

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

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

Chafe, W (1982) Integration and Involvement in Speaking, Writing, and Oral Literature, in Tannen, D (Ed.) *Spoken and Written Language: Exploring Orality and Literacy*, Norwood, New Jersey: Ablex Publishing Corporation, 35–53.

Channell, J (1994) *Vague Language*, Oxford: Oxford University Press.

Erman, B (1987) *A study of You Know, You See and I Mean in Face-to-face Conversation*, Stockholm: Minab Gotub.

Gilmore, A (2004) A comparison of textbook and authentic interactions, *English Language Teaching Journal*, 58/4, 363–374.

Halliday, M A K (1989) *Spoken and written language*, Oxford: Oxford University Press.

Leech, G, Rayson, P and Wilson, A (2001) *Word Frequencies in Written and Spoken English: based on the British National Corpus*, London: Longman.

McCarthy, M (1998) *Spoken Language and Applied Linguistics*, Cambridge: Cambridge University Press.

McCarthy, M and Carter, R (2003) What constitutes a basic spoken vocabulary? *Research Notes* 13, 5–7.

Marshall, H (2006) The Cambridge ESOL Item Banking system, *Research Notes* 23, 3–5.

Murray, S (2007) Broadening the cultural context of examination materials, *Research Notes* 27, 19–22.

O'Loughlin, K (1995) Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test, *Language Testing* 12, 217–237.

Rubin, J (1994) A Review of Second Language Listening Comprehension Research, *Modern Language Journal*, 78/2, 199–221.

Scott, M (1996) Comparing Corpora and identifying key words, collocations and frequency distributions through the WordSmith Tools suite of computer programmes, in Ghadessy, M, Henry, A and Roseberry, R L (Eds) *Small Corpus Studies and ELT*, Amsterdam/ Philadelphia: John Benjamins Publishing Company.

— (1999) WordSmith Tools Version 3.00.00, Oxford: Oxford University Press.

Stubbs, M (1986) Lexical density, a technique and some findings, in Coulthard (Ed.) *Talking about text*, University of Birmingham, Birmingham English Language Research.

— (1996) *Text and Corpus Analysis – computer assisted studies of language and culture*, Oxford: Blackwell.

Ure, J (1971) Lexical density and register difference, in Perrin, G E and Trim, J L (Eds), *Applications of Linguistics: Selected papers of the Second International Congress of Applied linguistics*, Cambridge: Cambridge University Press.

Using DIF to explore item difficulty in CAE Listening

ARDESHIR GERANPAYEH RESEARCH AND VALIDATION GROUP

Introduction

The issue of test fairness encompasses many concepts and models; chief among them is Differential Item Functioning. If test items operate in a differential fashion, then the scores for different groups are per se not comparable. Investigating differential item functioning (DIF) has long been practised by language test developers to demonstrate that tests are fair and relatively free from construct irrelevant variables (Alderman and Holland 1981, Chen and Henning 1985, Geranpayeh and Kunnan 2007, Kunnan 1990, and Ryan and Bachman 1992 to name a few)[1]. One of the emerging issues in recent investigations of Cambridge Main Suite examinations has been the shift in the traditional test population where test takers of many age groups, many of whom are unprepared, are sitting tests that are cognitively challenging. In such cases, it is the responsibility of the test development agency to educate the test users about the dangers of unprepared candidates

[1] The data reported here is the same as presented in Geranpayeh and Kunnan 2007.

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

taking such challenging tests and provide evidence to that effect. DIF analysis is a good means to investigate how multi-construct items can function differentially across different age groups.

Defining Differential Item Functioning

The AERA/APA/NCME standards (1999: 92–93) clearly emphasises the importance of '*equivalence*' and its relationship to candidates' background under *test interpretation* and use.

The AERA/APA/NCME standard 7.3 (1999: 81) states that

> 'When credible research reports that differential item functioning exists across age, gender, racial/ethnic, cultural, disability, and/or linguistic groups in the population of test takers in the content domain measured by the test, test developers should conduct appropriate studies when feasible. Such research should seek to detect and eliminate aspects of test design, content, and format that might bias test scores for particular groups.'

A test may be considered biased when it produces systematic differential performance among test takers from the same ability group but from different subgroups of interest (such as age, gender, race and ethnicity, nationality, academic major, native language, religion, and test takers with disability). Such systematic differential performance can be due to the presence of construct-irrelevant test characteristics in a test or in test items or relevant secondary item characteristics. Irrelevant item characteristics may be found in different components of a test: language variety, directions, content, response process, administration, and scoring, reporting, and interpretation towards a particular subgroup. Thus, differences in test performance for a designated subgroup of interest (DSI) could result in differences in meaning of test scores such that the validity of the test scores can be seriously in doubt. This definition also implies that test bias is not a result of random differential performance for a DSI when compared to another DSI as such comparisons would include ability levels. Therefore, a difference in performance mean for a DSI does not automatically mean that the test in question is biased.

Item bias procedures have received attention from the 1970s onwards because they were considered as a convenient procedure when no external criteria were available for such analysis. From the early stages, the focus was on the concept of relative item difficulty for different test taking groups. The idea was to match test takers with similar ability (as measured by the total score) from different subgroups. The expectation is that there would be comparable individual item difficulty for the subgroups as the test takers are matched in terms of overall ability. In cases where items performed or functioned differently for subgroups, such items were to be flagged and examined for potential bias. This procedure came to be known as Differential Item Functioning (Holland and Thayer 1988) and considerable literature (e.g. Holland and Wainer 1993) has developed around this concept and accompanying procedure. Some of the most common approaches to investigate DIF are:

- the Mantel-Haenszel statistic – Holland and Thayer (1988)
- the Standardisation procedures – Dorans and Kulick (1986)
- the Logistic Regression methods – Rogers and Swaminathan (1989), Zumbo (1999)
- the Logistic Discriminant function analysis
- Lord's Chi-Square – Lord (1977)
- Raju's Area measures – Raju (1988)
- the Marginal Maximum Likelihood Ratio test – Thissen, Steinberg and Wainer (1993).

The major benefit from this procedure, as Camilli and Shepard (1994:16) put it, has been to help 'clarify what a test is measuring and highlight the influence of irrelevant factors'. They cite the example of a study by Shepard et al. (1984) in which the researchers:

> '… found that verbal math problems were systematically more difficult for black examinees; that is, differences between blacks and whites were greater on this type of problem than on straight computations problems. Are verbal math problems biased against black? Not necessarily – solving verbal math problems is an important goal in its own right… However, once such results are known, they force re-examination and justification of the content framework of a test. In the case of the mathematics test with a heavy verbal competent, findings from the bias screening would prompt a more conscious appraisal of what proportion of the test items should be word problems and an effort to control the reading level required.' (Camilli and Shepard 1994:17)

This example is similar to the findings in Kunnan's (1990) study of the UCLA ESL placement test that showed DIF (in favour of test takers who were native speakers of Romance and Germanic languages) for vocabulary items that shared cognates with Romance and Germanic languages. Does this mean that such items are biased against test takers from other native language groups and should not be included in such a test? Not necessarily, but a deliberate appraisal of the proportion of such items in a vocabulary test needs to be conducted in order to respond to the concern that the test may have a disproportionately high number of cognates from Romance or Germanic languages.

Empirical studies

Table 1 presents a variety of studies in language testing that have focused on DIF since 1980 when test bias studies became the main approach to examine tests for fairness. As can be expected, most of the attention has been on differential performance with DIF methods taking the central role in such investigations. Investigations of tests focusing on DIF for test items with test takers from different native language backgrounds have been most popular. This may be due to the findings from second language learning studies that suggest when language learners study or test takers take a test in a second language, the main influence will be from their native language. Other test taker characteristics such as gender and academic major have also contributed to our understanding. While we have benefited from these studies, no clear and definitive findings regarding tests and test taking based on test taker characteristics have emerged as yet for language test development. Perhaps, this may be due to the general approach used by the researchers that is to detect whether test items displayed DIF but not to identify the causes of DIF.

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

Table 1: Empirical DIF studies in language testing (1980–2007)

Author and Year of Study	Specific focus
Swinton & Powers, 1980	Native language
Alderman & Holland, 1981	Native language
Shohamy, 1984	Test method
Alderson & Urquhart, 1985 a, b	Academic major
Chen & Henning, 1985	Native language
Zeidner, 1986, 1987	Gender, minorities
Hale, 1988	Major field & test content
Oltman et al., 1988	Native language
Kunnan, 1990	Native language, gender
Sasaki, 1991	Native language
Shohamy & Inbar, 1991	Question type & listening
Ryan & Bachman, 1992	Gender
Kunnan, 1992	Placement & exemption
Kunnan, 1995	Native language
Brown, 1993	Tape-mediated test
Ginther & Stevens, 1998	Native language, ethnicity
Norton & Stein, 1998	Text content
Brown, 1999	Native language
Takala & Kaftandjieva, 2000	Native language
Lowenberg, 2000	Different Englishes
Kim, 2001	Native language
Pae, 2004	Academic major
Uiterwijk & Vallen, 2005	Native language
Roever, 2007	Native Language

None of the studies cited examined the relationship between test performance and test takers' age. This is an important concern in the Cambridge ESOL Main Suite examinations, as there has been a shift in the traditional test population, where test takers of many age groups are taking these cognitively challenging tests. Anecdotal historical observations have indicated that if there were going to be any DIF in these examinations, it was likely to impact mainly listening items. Geranpayeh (2001) examined the country and age bias in the First Certificate in English (FCE) and recommended further investigation into DIF of listening item types. As there had been no empirical research in this area with the Certificate in Advanced English (CAE) examination, it was decided to investigate whether CAE Listening test items would exhibit DIF across age groups.

Research questions

Two research questions were formed in this study:

1. Do CAE listening paper test items exhibit DIF toward test taker groups in terms of age?
2. Are CAE listening paper test items biased toward test taker groups in terms of age?

To answer these questions, statistical analyses were performed first, items were then flagged for bias analysis, and content analyses were performed on these items as well as on the items that were not flagged.

Method

Data in this study are based on 4,941 test takers who took the Cambridge CAE examination in December 2002. Test takers' background information was collected through electronic Candidate Information Sheets (CIS) completed before the test administration. CIS included information about each candidate's gender, age, first language, years of study in English, and previous Cambridge exams taken. There were three versions of the CAE Listening paper, but our study only reports the performance of those who took Version 1. A total of 5,783 candidates sat for this paper, 25% of whom were Polish. To avoid the dominance of any particular cohort, we randomly removed half of the Polish candidates from the data, which reduced the total number of test takers to 4,941. Test takers were divided into three age groups: 17 and under, 18 to 22, and 23 and older. It was assumed that 18- to 22-year-old test takers are the target test takers for a CAE examination, which represent a range of test takers finishing high school to those studying at college. Seventeen-year-old and under test takers (called younger test takers) were assumed to be mainly high school students, whereas 23-year-old and older test takers were considered to be mature test takers. There were 83% of the test takers who fell into the first two age groups, which is also a typical representation of the CAE test taking population.

The test takers from this pool of 4,941 candidates were randomly reduced to 1,000 by BILOG-MG (Scientific Software International 2003) keeping in mind the proportion of test takers by age. This was done to make item estimation easier and to facilitate the interpretation of the significant differences found in any analysis. Large sample sizes tend to show statistically significant differences with minor variations in samples' performance, which in turn make the meaningfulness of the differences difficult to justify. Table 2 illustrates test takers' distribution by age of the sample.

Table 2: Distribution of candidates by age

Age	Total	Random Sample	% of Total
17 and under	1,871	411	41.1
18–22	2,247	422	42.2
23 and older	823	167	16.7
Total	4,941	1,000	100.0

N = 1,000

The test instrument

The Listening paper contains four parts. Each part contains a recorded text or texts and corresponding comprehension tasks. The texts in Parts 1, 3 and 4 are heard twice; the text in Part 2 is heard once only. The recordings contain a variety of accents corresponding to standard variants of English native speaker accent, and to English non-native speaker accents that approximate to the norms of native speaker accents. Background sounds are included before speaking begins, to provide contextual information. Subdued reaction from an audience to talks, speeches, etc., is also

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

included. For all parts of the paper candidates write their answers on an answer sheet. Each question in the paper carries one mark and the time allocated is approximately 45 minutes.

Analytical approaches

Two complementary approaches were used in this study: statistical analysis and content analysis. The statistical approach is discussed first.

The sample response data (N=1,000) from the CAE Listening paper was read into BILOG-MG (Scientific Software International 2003) for the first analysis. The Marginal Maximum Likelihood Ratio Test (Thissen, Steinberg and Wainer 1993) was used to investigate DIF that is said to be present when the probabilities of success on a given item are invariant between two or more groups at the same ability level. Moreover, we assume that DIF does not extend to the item discriminating powers. In other words, the b_j parameters for the separate groups are estimated on the assumption that the slope parameters, a_j, are homogeneous across groups (see Equation 1).

Equation 1: One-parameter logistic model

$$P_{(1)j}(\theta) = \frac{1}{1 + \exp[-a(\theta \ - b_j)]}$$

where,
$\exp(k) = e^k$ and
e = 2.718 is the base of the natural logarithm,
a is a scale constant determining the units of θ, and
b_j is a location parameter related to the difficulty of item j (also referred to as the item “threshold”). Items with larger values of b_j are more difficult; those with smaller values are easier.

We ran two different models: the compact model and the augmented model. In the compact model no group differences were assumed, whereas in the augmented model, we assumed that the items being investigated had DIF. We first tested the compact model, analysing the data in a single group as though they came from the same population, and calibrated the items accordingly. We noted the marginal maximum log likelihood of the item parameters in the final Newton cycles (labeled −2 LOG LIKELIHOOD in the output). We then analysed the data in separate groups using the augmented model, assuming the presence of DIF in the items and again noted the final −2 LOG LIKELIHOOD. Using a chi-square test (see Equation 2), we tested for the significance of difference between the two final −2 log likelihood estimates.

Equation 2: Chi-square test for the presence of DIF

$$G^2(df) = \frac{-2\log likelihood(C)}{-2\log likelihood(A)}$$

df = (n-1)(m-1),

where n is the number of items and m is the number of groups;

(C)=Compact model, (A)=Augmented model

When G^2 is significant, we may say that there is evidence that some items function differentially, rejecting the null hypothesis of no DIF effects on item locations.

The statistical analysis was followed by the content analysis. Following Roussos and Stout (2004) in terms of hypothesising the causes of DIF, the approach used here was to have the CAE subject officers to first examine the items that were identified as ones exhibiting DIF. If content analysis provided evidence that the DIF items were advantaging a particular group of test takers, then it might be possible to conclude that such items may be biased. An additional analysis included examining the remaining test items in order to identify items that might advantage a particular group of test takers although statistical analysis may not have identified them.

Results

Table 3 reports the descriptive statistics of raw scores by age. Average scores indicate that the groups are performing similarly on the test. However, the performance of the 17 and under age group appears to be slightly lower than the other two groups. Although the difference is only two and a half raw marks, it may indicate some adverse impact of the test for this particular group. However, most descriptive statistics indicators offer support to the view that the groups are performing relatively similarly on the test allowing us to conduct DIF analysis.

Table 3: Descriptive statistics of raw scores by age

	Age groups			
	17 & under	**18–22**	**23 & older**	**Average**
Mean	18.56	20.56	21.00	19.88
Std. Deviation	5.45	5.51	5.07	5.52
Median	19.00	21.00	21.00	20.00
Skewness	-0.06	-0.46	-0.29	-0.28
Kurtosis	-0.49	-0.21	-0.66	-0.46
Mean correct	58%	64%	66%	62%
Mean point Biserial correlation	0.36	0.38	0.35	0.37

Table 4 shows the item difficulty in the percent correct columns, and the logit, Pearson and Biserial values for all 32 items in the test. From all these indices, it can be observed that there is a range of item difficulty and item discrimination. Item 3 is the most difficult or least easy (0.37 on percent correct and 0.55 on logit) with only moderate item discrimination (biserial coefficient 0.40). Item 10 on the other hand is clearly the least difficult or easiest item (0.91 on percent correct and −2.36 on logit) but has lower item discrimination (point-biserial coefficient = 0.22).

Table 5 shows that there is significance difference between the two *−2 LOG LIKELIHOOD* in the models, indicating that the augmented (DIF) model better explained the data, i.e., there was evidence that differential item functioning was present.

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

Table 4: BILOG-MG item statistics output

Item Difficulty*				
Item	**% Correct**	**LOGIT**	**Item discrimination****	**Biserial**
1	0.6	-0.41	0.35	0.44
2	0.51	-0.02	0.33	0.41
3	0.37	0.55	0.32	0.4
4	0.56	-0.23	0.29	0.37
5	0.47	0.1	0.39	0.5
6	0.83	-1.61	0.29	0.43
7	0.52	-0.06	0.41	0.52
8	0.63	-0.52	0.43	0.55
9	0.78	-1.28	0.08	0.12
10	0.91	-2.36	0.22	0.39
11	0.44	0.22	0.37	0.46
12	0.6	-0.42	0.34	0.43
13	0.73	-1	0.24	0.32
14	0.78	-1.28	0.25	0.35
15	0.51	-0.05	0.29	0.36
16	0.44	0.26	0.34	0.43
17	0.44	0.26	0.22	0.28
18	0.5	0	0.16	0.2
19	0.5	0	0.18	0.23
20	0.75	-1.11	0.4	0.55
21	0.6	-0.42	0.34	0.43
22	0.68	-0.77	0.14	0.18
23	0.83	-1.59	0.35	0.52
24	0.71	-0.91	0.44	0.58
25	0.7	-0.85	0.22	0.29
26	0.73	-1.02	0.35	0.47
27	0.61	-0.44	0.42	0.54
28	0.89	-2.04	0.31	0.51
29	0.64	-0.59	0.42	0.55
30	0.45	0.2	0.09	0.11
31	0.55	-0.21	0.4	0.5
32	0.73	-1	0.4	0.54

Note. *N=1,000*

* % correct is based on Classical Test Theory and logit estimated by item response theory.

** Item discrimination index is based on the point-biserial Pearson correlations.

Table 5: Chi-square Test Results for Age Comparisons

	−2 LOG LIKELIHOOD
Compact	37,095.96
Augmented	36,886.08
G^2	209.88*

* Significant at $p<0.05$, df = 64

The software BILOG-MG produces a table for the group threshold differences for each item in the augmented model where Group 1 (age 17 and under) and Group 2 (age 18–22) are groups of candidates. Table 6 has two associated threshold difference figures in the two rightmost columns indicating the threshold differences between Groups 1, 2 and 3. Group 2 (age 18–22) was taken as the *reference* group as it comprises the biggest proportion of the candidates and is considered to be the target age group. These figures are immediately followed in the next row by the standard error figures for calculating the threshold differences. If the group threshold difference is bigger than two standard errors, we consider the threshold difference between the two groups to be significant at $p<0.05$. In other words, the item in question is functioning differentially between the groups and its content needs to be examined in the direction of the estimated contrasts in the b_j parameters. The standard errors are computed using Equation 3.

Equation 3:

$$s.e._{G2-G1} = \sqrt{\mathrm{var}(G2) + \mathrm{var}(G1)}$$

where G2, G1 are group of candidates, and var is their threshold variance.

Table 6 shows that of a possible 64 different comparisons only six items exhibited DIF: items 4, 11, 18, 20, 21 and 27. This is only three more than a random case where 3 items (5%) may have been found significant by chance alone. Of these six items, only Item 4 shows DIF on both group comparisons (1:2 and 2:3).

Discussion

Here we hypothesise the source of DIF for Item 4. Reviewing Table 4, Item 3, the immediate preceding item, was the most difficult of the items in the test. Since this is a listening test and the candidates had no control over the exposure of the items, it is quite possible that the candidates were still trying to respond to Item 3 when they were exposed to Item 4 and as a result they might have not listened to Item 4 in the most efficient way and hence might have missed part of the prompt clue. If that were the case, those who might have been affected by this factor would have certainly guessed the answer and their response should have little resemblance to the context of the response. There is some evidence to support this hypothesis. An analysis of the Common Wrong Answers[2] revealed that the three most frequent wrong answers were 'conversation group', 'consultation group' and 'conciliation group'. The correct answer was 'conservation group'. Were the younger candidates writing 'conversation group' – familiar to them from conversation classes in school? Were the older group writing 'consultation group' and 'conciliation group' – more sophisticated answers, although equally wrong? It is evident that all groups were having a listening problem and may have been trying to find an answer from their world experience of what it might be. In the light of the above discussion, it is possible to conclude that the significant difference in the threshold estimates for Item 4 was confounded by guessing and the candidates' world experience and it may not have related to the exhibition of DIF in this item.

It is difficult to comment on the other five items that exhibited DIF: items 11, 18, 20, 21 and 27. DIF only exists in one of the comparisons in these items. The DIF literature does not offer any consensus as how to deal with multiple group comparisons. Reise (personal communication, November 25, 2003) suggested combining the two groups that showed DIF in their comparison and evaluating them

[2] A Common Wrong Answer analysis is carried out in Cambridge ESOL for productive tasks in listening comprehension papers, where a sample of approximately 500 responses are captured during the live analysis to look for any possible acceptable response that might not have been conjectured at the time of test construction.

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

Table 6: Group Threshold Differences

	Group differences			Group differences	
Item	**Between Group 1 and 2**	**Between Group 2 and 3**	**Item**	**Between Group 1 and 2**	**Between Group 2 and 3**
1	-0.17 0.18	0.07 0.24	**17**	-0.42 0.17	0.30 0.23
2	-0.09 0.18	-0.05 0.24	**18**	-0.01 0.17	**0.57** 0.22
3	0.06 0.18	-0.20 0.23	**19**	-0.31 0.17	0.38 0.22
4	**-0.46** 0.18	**-0.61** 0.23	**20**	**0.55** 0.21	-0.07 0.30
5	0.11 0.18	-0.15 0.24	**21**	0.03 0.18	**0.61** 0.24
6	0.10 0.23	-0.32 0.35	**22**	-0.53 0.18	0.37 0.23
7	0.23 0.18	-0.17 0.24	**23**	0.00 0.23	-0.09 0.33
8	0.13 0.19	-0.11 0.27	**24**	-0.02 0.20	-0.40 0.29
9	-0.17 0.20	0.13 0.27	**25**	-0.17 0.18	0.02 0.25
10	-0.14 0.30	0.14 0.42	**26**	0.35 0.20	0.07 0.28
11	**0.48** 0.18	0.30 0.24	**27**	0.23 0.19	**-0.55** 0.27
12	0.18 0.18	0.05 0.24	**28**	0.12 0.27	-0.18 0.40
13	0.14 0.19	0.38 0.26	**29**	0.29 0.19	0.27 0.25
14	-0.25 0.20	-0.19 0.29	**30**	-0.37 0.16	-0.15 0.22
15	-0.12 0.17	0.07 0.23	**31**	0.30 0.18	-0.09 0.25
16	-0.33 0.18	0.08 0.25	**32**	0.27 0.20	-0.46 0.31

Note. Group 1 = (17 & under), Group 2 = (18–22), Group 3 = (23 & older); first row for each item is the threshold difference, second row for each time is the standard error of difference. Significant threshold differences are shown in bold.

against the third group. This cannot be meaningfully applied to our data because we would end up having two different threshold differences for the same items. If we applied such a methodology, we would lose the concept of the reference group. It would then become difficult to define what we meant by the 'younger' (17 and under) and 'older' (23 and older) groups.

A content analysis by the expert judges could not shed any further light on the source of the exhibited DIF in this study. The expert judges believed that the test items overall were suitable for the reference group 18 to 22, concluding that the items neither advantaged nor disadvantaged this group. They ruled out the presence of bias against the target age groups. This suggests that the CAE Listening test (Version 1 of the December 2002) is probably not biased against the test taker age groups included in this study. For full content analysis and relevant discussions see Geranpayeh and Kunnan (2007).

Conclusion

In this article we examined the Cambridge Certificate in Advanced English (CAE) test for DIF in terms of age. A two-step approach was used: first, the test items were examined for DIF and second, the items that were flagged were subject to content analysis through expert judges. The judges did not believe that the items were biased against specific age groups. No clear pattern emerged in terms of why the items were identified as exhibiting DIF. Further, the expert judges did not identify the causes of DIF for the items.

We hypothesised that the source of the invariance performance on Item 4 was related to test takers' difficulty in responding to Item 3 – its immediate preceding item. We did not, however, offer an explanation as to why the rest of the items exhibited

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

DIF. In the absence of any further empirical evidence, one possible explanation might lie in the different cognitive processes which test takers may have employed in attempting to answer the listening questions. It is possible that the ability to recall information or ability to use memory strategies may be critical in the items that exhibited DIF and different age groups might use these processes differently. This is difficult to investigate as the study did not collect data on test-taking strategies.

Finally, it is quite possible that the DIF exhibited could relate to the multidimensional nature of CAE listening items. Geranpayeh (2005a, 2005b) has recently shown that the CAE listening items have moderate to high correlations with items that test reading, writing and speaking skills in addition to having high correlations with items that test grammatical ability. In other words, the CAE listening items measure multiple dimensions to some extent. The large DIF values observed on some of the items are probably due to measuring those additional dimensions differently across the reference and the target groups. In the communicative approach to testing listening skills, on which the CAE is based, measuring secondary dimension is not only possible but also desirable and an intended part of CAE's focus. Testing the secondary dimension could be the source of variability in some of the items that exhibited DIF in this study. Further research could shed more light on this.

References and further reading

AERA/APA/NCME (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education) (1999) *Standards for educational and psychological testing*, Washington, DC: American Educational Research Association.

Alderman, D L and Holland, P W (1981) *Item performance across native language groups on the Test of English as a Foreign Language* (Research Rep. No. 81–16). Princeton, NJ: Educational Testing Service.

Alderson, J C and Urquhart, A (1985a) The effect of students' academic discipline on their performance on ESP reading tests, *Language Testing* 2, 192–204.

— (1985b) This test is unfair: I'm not an economist, in Hauptman, C, LeBlanc, R and Wesche M B (Eds), *Second language performance testing*, Ottawa, Canada: University of Ottawa Press, 15–24.

Brown, A (1993) The role of test taker feedback in the test development process: Test takers' reactions to a tape-mediated test of proficiency in spoken Japanese, *Language Testing* 10, 277–304.

Brown, J D (1999) The relative importance of persons, items, subtests and languages to TOEFL test variance, *Language Testing* 16, 217–238.

Camilli, G and Shepard, L (1994) *Methods for identifying biased test items*, Thousand Oaks, CA: Sage.

Chen, Z and Henning, G (1985) Linguistic and cultural bias in language proficiency tests, *Language Testing* 2/2, 155–163.

Dorans, N J and Kulick, E (1986) Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test, *Journal of Educational Measurement* 23, 355–368.

Geranpayeh, A (2001) *Country bias in FCE listening comprehension*, Cambridge ESOL Internal Research and Validation Report.

— (2005a) *Building the construct model for the CAE examination* Cambridge ESOL Internal Research and Validation Report.

— (2005b) *Language proficiency revisited: Demystifying the CAE construct*, paper presented at the 12th Language Testing Forum, Cambridge, England.

Geranpayeh, A and Kunnan, A J (2007) Differential Item Functioning in terms of age in the Certificate in Advanced English Examination, *Language Assessment Quarterly* 4/2, 190–222.

Ginther, A and Stevens, J (1998) Language background and ethnicity, and the internal construct of the Advanced Placement Spanish Language Examination, in Kunnan, A J (Ed.) *Validation in language assessment*, Cambridge: Cambridge University Press, 169–194.

Hale, G (1988) Student major field and text content: Interactive effects on reading comprehension in the TOEFL, *Language Testing* 5, 49–61.

Holland, P W and Wainer, H (1993) (Eds.) *Differential item functioning*, Hillsdale, NJ: Lawrence Erlbaum.

Holland, P W, and Thayer, D T (1988) Differential item performance and the Mantel-Haenzel procedure, in Wainer, H and Braun, H (Eds) *Test validity*, Hillsdale, NJ: Lawrence Erlbaum Associates, 129–145.

Kim, M (2001) Detecting DIF across the different language groups in a speaking test, *Language Testing* 18, 89–114.

Kunnan, A J (1990) Differential item functioning and native language and gender groups: The case of an ESL placement examination, *TESOL Quarterly* 24, 741–6.

— (1992) The author responds to comments on Kunnan (1990), *TESOL Quarterly* 26, 598–602.

— (1995) *Test taker characteristics and test performance*, Cambridge: Cambridge University Press.

Lord, F M (1977) A study of item bias, using item characteristic curve theory, in Poortinga, Y H (Ed.) *Basic problems in cross-cultural psychology*, Amsterdam: Swets and Zeitlinger, 19–29.

Lowenberg, P (2000) Non-native varieties and issues of fairness in testing English as a world language, in Kunnan, A J (Ed.) *Fairness and validation in language assessment*, Cambridge, Cambridge University Press, 43–59.

Norton, B and Stein, P (1998) Why the "Monkeys Passage" bombed: Tests, genres, and teaching, in Kunnan, A J (Ed.) *Validation in language assessment*, Mahwah, NJ: Lawrence Erlbaum, 231–249.

Oltman, P, Stricker, L and Barrows, T (1988) Native language, English proficiency and the structure of the TOEFL for several language groups (TOEFL Research Rep. No. 27). Princeton, NJ: Educational Testing Service.

Pae, T-I (2004) DIF for examinees with different academic backgrounds, *Language Testing* 21, 53–73.

Raju, N S (1988) The area between two item characteristic curves, *Psychometrika* 53, 495–502.

Reise, D (2003) Personal communication.

Roever, C (2007) DIF in the assessment of second language pragmatics, *Language Assessment Quarterly* 4/2, 165–189.

Rogers, J and Swaminathan, H (1989) *A logistic regression procedure for detecting item bias*, paper presented at the annual meeting of the American Educational Research Association, San Francisco, March.

Roussos, L and Stout, W (1996) A multidimensionality-based DIF analysis paradigm, *Applied Measurement in Education* 20, 355–371.

Ryan, K E and Bachman, L F (1992) Differential item functioning on two tests of EFL proficiency, *Language Testing* 9/1, 12–29.

Sasaki, M (1991) A comparison of two methods for detecting DIF in an ESL placement test, *Language Testing* 8, 95–111.

Shepard, L, Camilli, G and Williams, D M (1984) Accounting for statistical artifacts in item bias research, *Journal of Educational and Behavioral Statistics*, 9/2, 93–128.

Shohamy, E (1984) Does the testing method make a difference? The case of reading comprehension, *Language Testing* 1, 147–170.

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

Shohamy, E and Inbar, O (1991) Validation of listening comprehension tests: the effect of text and question type, *Language Testing* 8, 41–66.

Swinton, S and Powers, D (1980) Factor analysis of the TOEFL for several language groups (TOEFL Research Rep. No. 6), Princeton, NJ: Educational Testing Service.

Takala, S and Kaftandjieva, F (2000) Test fairness: A DIF analysis of an L2 vocabulary test, *Language Testing* 17, 323–340.

Thissen, D, Steinberg, L and Wainer, H (1993) Detection of differential item functioning using the parameters of item response models, in Holland, P W and Wainer, H (Eds) *Differential item functioning,* Hillsdale, NJ: Lawrence Erlbaum, 67–113.

Uiterwijk, H and Vallen, T (2005) Linguistic sources of item bias for second-generation immigrants in Dutch tests, *Language Testing* 22, 211–234.

Zeidner, M (1986) Are English language aptitude tests biased towards culturally different minority groups? Some Israeli findings, *Language Testing* 3, 80–95.

– (1987) A comparison of ethnic, sex, and age biases in the predictive validity of English language aptitude tests: Some Israeli data, *Language Testing* 4, 55–71.

Zumbo, B (1999) *A Handbook on the theory and methods of differential item functioning: logistic regression modelling as a unitary framework for binary and Likert-Type (ordinal) item scores*, Ottawa: Directorate of Human Resources Research and Evaluation, Department of National Defence.

Adapting listening tests for on-screen use

ED HACKETT ASSESSMENT AND OPERATIONS GROUP

Introduction

Cambridge ESOL has produced listening tests for computer-based exams since 1999. Prior to 2005, these tests were CD-ROM based, but in November 2005, Cambridge ESOL launched its first Internet delivered computer-based exam on the *Cambridge Connect* system. Ahead of the launch of the computer-based Cambridge Preliminary English Test (CB PET), extensive development and trialling was carried out to re-evaluate the format of listening tests for computer-based tests as compared to their paper-based variants.

There are a number of key issues to consider in adapting paper-based tests for on-screen use: display of the items, the mechanics involved in responding to questions, adaptation of the rubrics to cater for any changes, and how the candidate interacts with the audio. The impact of these changes then has to be evaluated to see what possible effect this may have on ease of use and comparability of the two different forms.

This article highlights some of the issues in adapting listening tests for on-screen use and discusses recent developments made as the range of products using Cambridge Connect has expanded over the past two years.

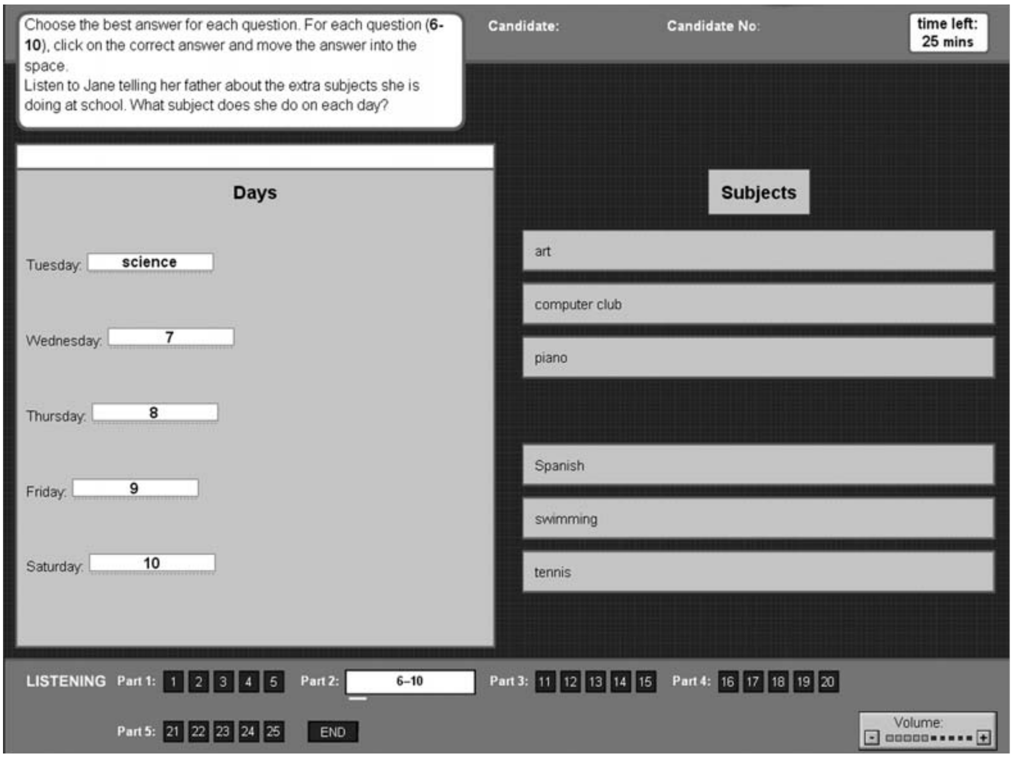
Screen design and answer mode

One key difference in adapting a PB product for on-screen use is screen aspect. Hackett (2005) notes the impact of screen orientation, with most PB tests displaying in portrait view, whereas most computer screens display in landscape format. This inevitably impacts on the location of questions, text and visuals on the page. Unlike Reading, where scrolling was essential for many multi-option tasks, the majority of Listening tasks could fit on a single screen. CB PET, with no task having more than 6 items, was able to display this on a full screen. However, CB BEC Vantage, with one task containing 8 items, had to employ scrolling for Items 7 and 8. However, as the items in this task have a linear relationship with the audio recording, i.e. Items 7 and 8 relate to the last part of the text; candidates only have to scroll down at one point in the task.

Answer mode is another key difference between CB and PB tests. Cambridge ESOL General English and Business English tests allow candidates to mark their answers on the question paper before transferring them onto a scannable answer sheet at the end of the test – additional time is allowed for this. For CB tests, the candidate usually clicks on a button next to the item they have chosen or types text in a given space. CB BULATS, introduced in 1999, was one of the earlier CD-ROM computer-based tests produced by Cambridge ESOL and only contained multiple-choice answers, so a simple radio button answer format was employed. The use of radio buttons for CB BULATS did not raise any issues as, being an adaptive test; candidates had to complete each task before another one was selected for them from the item bank. CB PET, launched in 2005, introduced additional needs to CB BULATS as regards the mode of answering. Firstly, being a linear test, modelling the PB format, candidates had the choice of returning to earlier unanswered questions. Radio buttons once selected can be changed, e.g. from A to B or C, but cannot be deselected, i.e. shown as unanswered as a flag for the candidate to return to later. Early trials highlighted this as a highly desirable function for candidates taking CB PET. The ability to select and then deselect an answer so that it could be returned to later was added to the answer button functionality for all CB tests on the Connect platform.

The need to display productive, note filling, tasks on screen was another element that required further development of the CB format. The ability to write as you listen is taken as a given skill in PB tests, but the ability to type as you listen is not necessarily a universal skill for all candidates, even in their native language. Early trials for CB PET showed the need to factor in transfer time for candidates not comfortable with typing as they listen. Candidates are permitted to take notes as they listen, then type in the answer, with a short period of time being allowed at the end of a task for this. In trials (Hackett ibid.),

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.



Choose the best answer for each question. For each question (**6-10**), click on the correct answer and move the answer into the space.
Listen to Jane telling her father about the extra subjects she is doing at school. What subject does she do on each day?

Candidate:

Candidate No:

time left: 25 mins

Days

Tuesday: science

Wednesday: 7

Thursday: 8

Friday: 9

Saturday: 10

Subjects

art

computer club

piano

Spanish

swimming

tennis

Figure 1: CB KET Listening Part 2 – Drag and Drop answer format

53% of respondents said that they made notes then typed their answers, whilst 42% indicated that they typed as they listened. The remaining percentage expressed no preference. Given increasing computer familiarity, especially amongst candidates of school age, it would be useful to revisit this statistic in the future.

CB KET and CB BEC Preliminary and Vantage, launched in 2007, presented the opportunity for additional functionality, with drag and drop being introduced for tasks with matching exercises (see Figure 1). As with earlier innovations, these changes created no identified problems for the candidature in trialling carried out in 2007.

In addition to these functional changes, the adaptation of paper-based tests to CB format also allowed scope for the enhancement of images which, in addition to being in colour, used vector graphics. Vector graphics work on a geometric model so they can adapt to different screen sizes and shapes without the distortion or degradation of quality that can be associated with raster graphics, e.g. jpegs, which use pixels to represent an image.

Adaptation of soundtrack and rubrics

Adaptations to rubrics were minor, with most changes being necessitated by the change of answer mode, e.g. *click* rather than *tick* or *mark*, and *type* rather than *write*. And, as mentioned above, a small amount of additional time was added to allow for typed answers in note-taking exercises. Minor pauses were also added to ensure that candidates had selected the appropriate task and question, e.g. '*Part One, question two*', is followed by a pause of 2 seconds before the audio for this question begins. If a candidate momentarily loses their place on a question paper, a quick turn of the page usually suffices, but in CB format, they might need a fraction longer to navigate to the correct question. These additional timings were more than offset by the elimination of time at the end of a PB test for transferring answers to an answer sheet, 6 minutes being allowed for this at the end of the test in PB PET.

Aside from these changes, the test audio is the same as that for the PB equivalent, both formats being in lockstep linear mode, i.e. the candidate answers the questions in the same order as the answers are presented in the Listening text. The sound audio for Connect is in MP3 format as opposed to CD audio and, as test bundles are downloaded to a control PC prior to the start of the test (Seddon 2005), streaming over the internet, and any problems associated with this, are avoided. Candidates listen to the audio on headphones and can adjust the volume to suit personal preferences, but this functionality is also available for some PB administered listening tests.

Comparability of CB and PB Listening

Whilst the majority of candidates taking Cambridge ESOL listening tests on computer appear to prefer this format to the equivalent PB variant (87% of CB PET trial candidates expressed a preference for CB, Hackett ibid.), the question of comparability inevitably arises. Though the task formats for Cambridge ESOL exams on Connect is the same for both PB and CB versions, it would be difficult to claim direct equivalence, i.e. that a candidate would score the same on both formats. As with a candidate taking two PB forms of a test, there will inevitably be some minor variation in performance and result. However, what we hope to see is comparability; that is that groups of candidates from similar backgrounds and abilities achieve similar scores on both forms of the test. Whilst CB PET has been running for over two years, the size of the candidature and the relatively small number of countries participating in online testing to

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

date means it is difficult to draw direct comparisons. However, despite this differing candidature, CB results are not greatly different to PB, with similar pass rates. The average pass rate for PB PET sessions in 2007 was 76.23% and the average pass rate for CB PET up to and including February 2008 was 76.1%. Earlier research into CB/PB comparability on BULATS (Jones 2000) and IELTS (Blackhurst 2005, Green and Maycock 2004) also found no significant differences between PB and CB performance and concluded that both forms of the test could be used interchangeably.

It is not surprising that CB PET has proved popular with candidates taking this form of the exam as the candidature is predominately made up of people who have grown up with computers and technology, 77% of the candidates are aged 18 or under. The number of centres approved for running Cambridge ESOL examinations online has grown rapidly over the past year, with nearly 100 centres in 27 countries approved to offer one or more Cambridge exams on the Connect platform. As this number continues to grow, we will be able to gather further data on the comparability of paper-based and computer-based exams.

References and further reading

Blackhurst, A (2005) Listening, Reading and Writing on computer-based and paper-based versions of IELTS, *Research Notes* 21, 14–17.

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

Hackett, E (2005) The development of a computer-based version of PET, *Research Notes* 22, 9–13.

Jones, N (2000) BULATS: A case study comparing computer-based and paper-and-pencil tests, *Research Notes* 3, 10–13.

Seddon, P (2005) An overview of computer-based testing, *Research Notes* 22, 8–9.

Recent publications of interest

Studies in Language Testing

January 2008 saw the publication of another title in the *Studies in Language Testing* series, published jointly by Cambridge ESOL and Cambridge University Press. Volume 23 in the series, by Professor Alan Davies, is entitled *Assessing Academic English: Testing English proficiency, 1950–1989 – the IELTS solution*.

This latest volume presents an authoritative account of how academic language proficiency testing evolved in the UK, and later Australia. It chronicles the early development and use of the English Proficiency Test Battery (EPTB) in the 1960s, followed by the creation and implementation of the revolutionary English Language Testing Service (ELTS) in the 1970s and 1980s, and the introduction of the International English Language Testing System (IELTS) in 1989. The book offers a coherent socio-cultural analysis of the changes in language testing and an explanation of why history matters as much in this field as elsewhere. It discusses the significant factors impacting on language test design, development, implementation and revision and presents historical documents relating to the language tests discussed in the volume, including facsimile copies of original test versions, such as the first versions of ELTS in 1980 and of IELTS in 1989. The volume will be of considerable value to language test developers and policy-makers, as well as teachers, lecturers and researchers interested in assessing English for Academic Purposes (EAP) and in the role played by ELTS and IELTS over the past 25 years or more. More information is available at: www.cambridgeesol.org/what-we-do/research/silt.html

Mike Milanovic and Alan Davies at the launch of Studies in Language Testing 23: Assessing Academic English at AAAL 2008.

There were book launches and author signings at the AAAL 2008 Annual Conference in Washington DC in March and the TESOL 2008 conference in New York in early April[1]. Conference delegates had the opportunity to meet both Alan Davies and SiLT Series Editor Mike Milanovic at these events, pictured below.

Publications by ESOL research staff

The beginning of 2008 also saw publication by Blackwell of the new *Handbook of Educational Linguistics* edited by Bernard Spolsky (formerly Bar Ilan University) and Francis M Hult (University of Texas at San Antonio). This new Handbook is described as a dynamic, scientifically grounded overview which reveals the complexity of this growing field while remaining accessible for students, researchers, language educators, curriculum developers, and educational policy makers. It takes into account the diverse theoretical foundations, core themes, major findings, and practical

[1] See page 30 for a review of a Cambridge ESOL jointly-led symposium at AAAL 2008.

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

applications of educational linguistics. The section on core themes includes a contribution by Neil Jones and Nick Saville on the topic of *Scales and Frameworks*.

Two other members of the ESOL Research and Validation Group have recently contributed chapters based upon Cambridge ESOL's ongoing research as follows: Andrew Blackhurst has a paper entitled *Computer-based and Paper-based Versions of IELTS* in Alexander, O (2008) (Ed.) *New Approaches to Materials Development for Language Learning: Proceedings of the 2005 joint BALEAP/SATEFL conference*; Karen Ashton has a paper entitled *The Languages Ladder and Asset Languages: A New Assessment Framework for Languages in England* in Kenner, C and Hickey, T M (in press) (Eds) *Multilingual Learning Communities across Europe* (to appear later in 2008).

ESOL staff seminar programme 2007–8

The Research and Validation Group coordinate an annual series of monthly seminars and workshops on a range of language testing and related topics for Cambridge ESOL staff and specially invited colleagues. The seminars which took place in 2007 and early 2008 represent a cross-section of the work and concerns of Cambridge ESOL and are summarised below.

Issues in Testing English for Specific Purposes

In January 2007, Martin Robinson and David Thighe spoke about issues in testing English for Specific Purposes (ESP). Testing English for Specific Purposes, such as English in a legal or finance-related work setting, is becoming increasingly important for Cambridge ESOL. Recently, a number of tests in ESP have been developed by Cambridge ESOL, including The International Legal English Certificate (ILEC) and The International Certificate in Financial English (ICFE). David Thighe (Research and Validation Group) gave an overview of the particular issues related to developing tests for ESP, such as current definitions of ESP tests through the notion of authenticity, the inseparability of background or content knowledge from language knowledge, and the need to ascertain the degree of specificity of tests. Next, Martin Robinson (Test Development Unit) presented on the role of content knowledge specialists, e.g. lawyers in relation to ILEC, in the test production cycle, and argued that content specialists play a crucial role in the operational production of ESP tests through their insights into the target language use situation.

Deconstructing the Main Suite tests to understand them better

Stuart Shaw and Hanan Khalifa presented in February on the exams which make up our Main Suite product group, i.e. KET, PET, FCE, CAE and CPE. They explained and discussed the three key factors in any language test – the test taker's cognitive abilities, features of task and context, and the scoring process – to show how these factors form a triangular relationship at the heart of any assessment activity. Drawing on recent analyses of the Writing and Reading components for our Main Suite exams which contributed to SiLT volumes on the constructs of second language writing and reading ability (Shaw and Weir 2007, Khalifa and Weir in prep), they illustrated how certain elements of each factor can be manipulated in order to clearly differentiate one proficiency level from another. Being able to demonstrate a clear understanding of how we conceptualise language proficiency, in terms of underlying abilities or construct(s), and how we operationalise these constructs for assessment purposes in our ESOL exams, is important to enable us to support the claims we make about the usefulness of our tests.

Test impact: where next?

In March Nick Saville and Roger Hawkey (an ESOL consultant) reported on impact studies and the role Cambridge ESOL has played in developing this methodology. Nick Saville was already at UCLES when they first institutionalised the study of test impact. In fact, Mike Milanovic and he take credit for putting impact into context along with the V, the R and the P of test VRIP (validity, reliability, impact and practicality). Nick began this presentation-and-workshop session with his recollections of why VRIP came to the fore when it did, why research into impact was seen as a vital part of UCLES EFL/Cambridge ESOL credibility as a major international test provider.

Roger Hawkey then discussed recent Cambridge ESOL impact studies which he co-ordinated. These include studies on IELTS, on the Italian *Progetto Lingue 2000* language teaching reform project, on *CPE* textbook washback, and the beginnings of a new study, of the Cambridge CRUI online blended learning course for B1 qualification. Roger Hawkey's experiences seeking data on the consequential validity of ESOL tests as they affect a whole range of our stakeholders provided useful insights into impact study design, instrument development, data collection and findings.

Finally, Nick Saville discussed impact studies in current Cambridge ESOL thinking and action and predicted the role and status of impact studies in *future* ESOL policy, strategies and structures.

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

Paired interaction

April's seminar considered paired spoken interaction and was led by Lynda Taylor. A key differentiating feature of our examinations is that they include direct assessment of speaking and writing. This can be traced back to the early days of CPE in 1913 and reflects a view of linguistic proficiency that embraces ability to use language for communication as well as knowledge about language. Over recent decades, the advent and spread of communicative teaching methods have influenced the shape and content of language tests even more strongly.

In the first part of this seminar, Lynda Taylor looked at how Cambridge ESOL exams test spoken language proficiency. Many of our tests now use a paired format for the speaking component, in which two candidates interact with an examiner and undertake a series of tasks. Lynda also considered the pairing of raters, considering questions such as: Why do we have two raters as well as two candidates in many of our speaking tests? What are their roles?

English Profile

In May, Svetlana Kurtes, Nick Saville and two research students (Caroline Williams and Oestein Andersen) presented on English Profile. English Profile is a long term collaborative programme of research, consultation and publication, designed to enhance the learning, teaching and assessment of English worldwide. Building on existing resources such as the Common European Framework of Reference for Languages, and the Breakthrough, Waystage, Threshold and Vantage specifications, a multi-disciplinary team is working to produce Reference Level Descriptions for English. These will provide a uniquely detailed and objective analysis of what levels of achievement in language learning actually mean in terms of the grammar, vocabulary and discourse features that learners can be expected to have mastered.

At the heart of English Profile there is an extensive research programme which involves analysing Cambridge ESOL's Cambridge Learner Corpus (CLC), focusing on the specific errors made by learners from different linguistic backgrounds. This seminar provided an overview of the work the research team is doing and explained its relevance to the project as a whole. It described the role of new tagging, parsing and indexing tools that enable more sophisticated analysis of CLC content. Caroline Williams and Oestein Andersen, both PhD students at Cambridge University, presented their current work looking at patterns of lexical choice errors by level and first language, and how these can be detected automatically.

Cambridge ESOL in Latin America

June's seminar was led by Sharon Harvey who considered Cambridge ESOL's activities in Latin America. Cambridge ESOL has been offering examination services in Latin America since the early 20th Century. Traditionally our exams were taken by students studying in a small number of private language institutes; nowadays our reach extends to providing language assessment to more than half a million Colombian school leavers, in addition to providing the full range of Cambridge ESOL examinations across the continent. This seminar discussed the potential and the challenges of the Latin American markets and examined how Cambridge ESOL has continued to develop the markets for mature exams such as FCE, how we have successfully launched new products such as TKT and CB PET, and how we have positioned ourselves as language testing experts who are increasingly sought to assist national and regional governments with their language development policies.

Getting closer to our Stakeholders

In July, a group of 7 staff from across Cambridge ESOL presented on the relationships that we have with our diverse stakeholders. The range of stakeholders that Cambridge ESOL deals with has broadened significantly over the last few years, such that there is now a far more complex interaction between different stakeholders in the testing process. These relationships are managed by a range of people in a number of different groups in Cambridge ESOL, and the speakers represented four departments: Communications and Stakeholder Relations, Assessment and Operations Group, Customer Services Group, and Business Development and Business Management Group.

Jenny Grewcock (Communications and Stakeholder Relations) began the seminar by providing an overview of our stakeholder community, how we define stakeholders and our approach to stakeholder relations. Juliet Wilson (Customer Services Group) then spoke about the ways in which the Customer Services Group aims to serve our centres as best we can. This Group consists of three main areas: the Cambridge ESOL Helpdesk (centres' first point of contact for any queries), Application Support (the team that provides technical training to centres and technical support of systems and products), and the Centre Management Unit (including centre registration, centre support, centre inspections and the London Open Centre). Next, Debbie Howden (Business Development and Business Management Group) reported on the Centre Consultation Survey, and its uses for finding out how we can improve our service to centres.

Nic Underhill (Assessment and Operations Group) presented on the Professional Support Network (PSN) of external consultant resources (such as oral and writing examiners, team leaders, seminar presenters, centre inspectors) and some of the systems that are being developed to better support the relationships we have with this vital network. Simon Fenn (Customer Services Group) talked about the redevelopment of our current website, so that we can serve our various stakeholder communities in a more dynamic, functional and relevant way; our main site will now segment information for different stakeholder groups by providing portals to other Cambridge ESOL sites. Then Mickey Bonin (Business Development and Business Management Group) provided a business development perspective of the kind of stakeholders this group deals with, and the kind of projects Business Development Managers are involved in. Andrew Nye (Communications

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

and Stakeholder Relations) concluded the seminar by talking through the Stakeholder Relationship Questionnaire – a recent initiative to find out whether our key stakeholders believe we are strengthening and improving our relationships in the ways that we think we are.

Grammar and spoken language

There was a break over the summer until September's session when we invited Professor Mike McCarthy to speak on the topic of grammar and spoken language. Thanks to the availability of spoken and written corpora, many new insights have been gained about contemporary English grammar as it is used by a wide range of people of different ages and social and geographical backgrounds. Corpus analysis reveals that the core of the language includes items with important interpersonal grammatical functions, as well as the traditional categories of tense, number, etc. Most significantly, corpora enable use to observe differences between written and spoken grammar, including grammatical items and patterns not previously noticed or recorded, things often considered 'wrong' or 'bad English' but common in the speech of educated users and items and patterns only occurring in particular contexts or genres. Mike offered examples from spoken British and American English corpora and argue that spoken grammar displays key characteristics which require a different approach to description. In an era when variety and diversity are being stressed in language education, grammar should be no exception.

Quality Assurance – ISO and ALTE

October's seminar was led by Dittany Rose and Michael Corrigan who spoke about quality assurance in two areas of Cambridge ESOL's work. Quality assurance has long been an important part of what we do at Cambridge ESOL but audits are a relatively new feature to the quality management system. This seminar explored the relationship between audits and wider quality management concerns and indicated what this means for Cambridge ESOL as a whole and for individual members of staff. In the first half of this seminar, Dittany Rose spoke about Cambridge ESOL internal audits and the quality management system. In the second part, Michael Corrigan covered ALTE audits, which were designed to apply specifically to language testing. Cambridge ESOL is involved in both types of audit and the seminar raised awareness for staff throughout Cambridge ESOL, some of whom are involved in the quality assurance process as internal auditors.

Using corpora for language assessment

In November, Fiona Barker and Svetlana Kurtes led a seminar on the use of corpora for language assessment, including its development over the last few decades, major contributions to the field, work in progress and some of the challenges faced by language testers, culminating with what Cambridge ESOL is doing in this area. Current and future directions for using corpora in language testing were presented, including developing and researching our own corpus resources within English Profile – a collaborative research endeavour to develop reference level descriptors for English.

This session also updated colleagues on our own Cambridge Learner Corpus, which includes candidate scripts and question papers, and our growing collection of speaking tests. Colleagues were encouraged to read Fiona Barker and Lynda Taylor's chapter with the same title in the 7th volume of the 2nd edition of the *Encyclopedia of Language and Education* (Taylor and Barker 2008).

21st century perspectives on the Specific Purpose Language construct

In December, we welcomed two external speakers to Cambridge.

In the first December seminar, Professor Dan Douglas (Iowa State University, USA) spoke on testing Language for Specific Purposes. The nature of the construct underlying specific purpose language courses and tests has been debated for nearly a quarter of a century, since Widdowson's (1983) investigation of the rationale for the specific purpose language enterprise. Douglas (2000) has argued that a definition of specific purpose language must focus on an interaction between language knowledge and specific purpose background knowledge and that background knowledge, far from being a factor leading to 'construct irrelevant variance' (Messick 1989:35), is essential to defining the construct. This entails a willingness, indeed necessity, to include non-linguistic elements in the LSP construct definition. Chapelle (1998), arguing for an 'interactionalist perspective' on construct definition, points out that our definitions of language ability must change with contexts of use while our understanding of contexts must be influenced by the language associated with them. Taking this notion a step further by considering the place of the technological means of communication, Chapelle and Douglas (2006) define language ability as the ability to select and deploy appropriate language through the technologies that are appropriate for a situation. In this seminar, Dan Douglas explored ways in which context, particularly mediating technology, and language ability interact in defining the construct of specific purpose language ability.

Social cohesion and language learning

In the second December seminar, Professor Joseph LoBianco (University of Melbourne) spoke on the relationship between social cohesion and language learning. He discussed the term 'social cohesion' and its connection with languages against the backdrop of two key considerations. The first of which is the dramatic transformation of human societies worldwide under conditions of globalisation in which population is central and at unprecedented levels. Combined with demographic shifts in fertility rates these changes appear to be decisive and permanent and produce multicultural societies

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

everywhere. Contrasting with this is the play of both nostalgia and established practices of states and education systems which are premised on uniformity. At the supranational level we see both: the instrumental rationality of efficiency, which dominates in economics and regional security but which is contested by the stubborn resistance of tradition and even atavism. The traditional aspiration of most national states has been for linguistic uniformity and the desire for secure homelands recognisable by the cultural continuity and tradition. In reality this is often a myth, but a myth on which many national states have been forged, community imagined and economies constructed. The talk commented on the contribution of research in language education in forging broad, citizenship based, socio-cultural community and cohesion and included examples of language education practice drawn from examples worldwide to underscore the links between practical work in educational institutions with high level policy ambition and national aspiration for social cohesion and security.

Cambridge ESOL's approach to test fairness

In January 2008, Lynda Taylor and Ardeshir Geranpayeh considered test fairness alongside related matters such as test washback and impact, test standards and test bias, equity and ethics for testing professionals as well as maintaining a concern for the technical qualities of tests, such as reliability and validity.

In the first part of this session, Lynda Taylor briefly reviewed the growth of interest in test fairness and the issues under current debate within the language assessment community. She considered the Cambridge ESOL view on test fairness and examined the measures the organisation undertakes in its efforts to achieve fair outcomes for all test stakeholders. This part of the seminar provided some helpful answers to the frequently asked question 'How do we know Cambridge ESOL exams are fair?'

In the second part, Ardeshir Geranpayeh provided an example of how Cambridge ESOL continuously monitors their tests in terms of whether test takers are receiving a fair test. He reported on Differential Item Functioning (DIF) in terms of age in the Certificate in Advanced English (CAE) examination (see his article in this issue). He explained how he and Antony Kunnan investigated whether the test items on the Listening section of CAE functioned differently for test takers from three different age groups (Geranpayeh and Kunnan 2007).

Language, migration and social integration: Implications for assessment

In February, an external speaker Dr Philida Schellekens (independent consultant) co-presented with a Cambridge ESOL Research and Validation colleague on the implications of recent government policies on language assessment in general, and the work of Cambridge ESOL in particular.

Philida Schellekens talked about the wider context of government policy since 2001, when the government in England launched its strategy to improve the literacy and numeracy skills of its population. She showed how the Skills for Life strategy has become a major government initiative which has been generously funded and energetically pursued. National standards for literacy and numeracy have been created and targets have been set to monitor delivery over time. She then went on to discuss how well the strategy has fared six years later. The audience reflected on what can be learnt from the experience, both in terms of expected and unexpected consequences. We discussed, for example, the consequences of the general practice that the achievement of migrants and refugees who need English for social and work purposes is assessed against the national literacy standards, which were designed for native English speakers. This has had important implications for curriculum design, testing and classroom delivery. Secondly, Philida explained that government departments have used the document *Pathways to Proficiency* to align EFL and ESOL qualifications but showed that there are doubts about the accuracy of these calibrations.

In the second half of this seminar, Szilvia Papp gave a brief historical background to language testing for citizenship and migration purposes since 2002. She then reviewed the work of the Language Assessment for Migration and Integration (LAMI) subgroup within ALTE. She briefly examined the publicly available materials provided for study towards the UK citizenship test: the *Life in the United Kingdom: A Journey to Citizenship* handbook (The Home Office 2007). The audience discussed to what extent and how the language used in these materials reflects the targeted level of proficiency (minimum Entry Level 3 within the National Qualifications Framework, or B1 in the CEFR) and language use domain (functional competence required for successful demonstration of citizenship or residency).

References

Chapelle, C (1998) Construct definition and validity inquiry in SLA research, in Bachman, L and Cohen, A (Eds), *Interfaces Between Second Language Acquisition and Language Testing Research*, Cambridge: Cambridge University Press, 32–70.

Chapelle, C and Douglas, D (2006) *Assessing Languages through Computer Technology*, Cambridge: Cambridge University Press.

Douglas, D (2000) *Assessing Languages for Specific Purposes*, Cambridge: Cambridge University Press.

Geranpayeh, A and Kunnan, A (2007) Differential Item Functioning in Terms of Age in the Certificate in Advanced English Examination, *Language Assessment Quarterly*, 4/2, 190–222.

Khalifa, H and Weir, C J (in preparation) *Examining Second Language Reading*, Studies in Language Testing, Cambridge: UCLES/Cambridge University Press.

Messick, S (1989) Validity, in Linn, R L (Ed.) *Educational Measurement*, New York: Macmillan, 13–23.

Shaw, S D and Weir, C J (2007) *Examining Writing: research and practice in assessing second Language writing*, Studies in Language Testing volume 26, Cambridge: UCLES/Cambridge University Press.

Taylor, L and Barker, F (2008) Using Corpora for Language Assessment, in Shohamy, E and Hornberger, N (Eds) *Encyclopedia of Language and Education* (2nd Ed.), volume 7, New York: Springer, 241–254.

Widdowson, H (1983) *Learning Purpose and Language Use*, Oxford: Oxford University Press.

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

Conference reports

Cambridge ESOL staff have recently taken part in a number of key international events, reported on below.

Cambridge ESOL and CEFR familiarisation activities

As part of its ongoing relationship with the Council of Europe CEFR levels, Cambridge ESOL weaves into its test development and operation cycle a variety of activities aimed at familiarising its staff and consultants – especially those who are involved in test development, item writing or rating scale development and implementation – with the Common European Framework. We report here on one of the recent familiarisation activities which took place in November 2007.

The intended outcomes of the familiarisation activity were as follows:

1. a common understanding of the aims and aspirations of the CEFR and its descriptive scheme
2. a shared knowledge of differentiating features across certain level thresholds, i.e. B1/B2 and B2/C1
3. a participant-led action plan for cascading new knowledge, skills and attitudes gained as a result of this familiarisation activity.

The activity took the form of a one-day workshop with pre- and post workshop tasks. The workshop was led by one of Cambridge ESOL consultants Dr. Lynda Taylor and attended by a mixture of subject officers, validation officers, item writer chairs, senior team leaders, and principal examiners.

Pre-workshop tasks aimed at raising participants' awareness to how the development of the CEFR and its associated projects e.g. the European Language Portfolio has affected the development of Cambridge ESOL examinations; and at encouraging participants to reflect on how the use of the CEFR has affected their own work on Cambridge ESOL examinations, e.g., the work of item writers, local examiners, regional team leaders, assistant subject officers, examination administrators, etc. Other pre-workshop tasks aimed at ensuring common understanding of the CEFR global scale and selected B1 to C1 descriptors related to the four language skills: listening, speaking, reading and writing. A classification exercise was used to achieve this aim. A further task involved using the CEFR global scale to self-assess participants' own ability in a second language. It is worth mentioning here that the pre-workshop tasks are similar to those recommended by the Manual for relating language examinations to the CEFR (preliminary pilot version 2003).

The face-to-face workshop itself started with an introductory focus on the origins, aims and nature of the CEFR, its relevance for language assessment and finally its relevance and implications for participants as professional language testers working with Cambridge ESOL. The workshop then moved on to a descriptor-sorting activity, thus building on one of the pre-workshop tasks. Group discussions were recorded to be analysed at a later stage in order to examine the rationale and justification for assigning each descriptor to its level. The workshop ended by training participants in applying skill-specific CEFR B1 to C1 level scales to actual tasks and performances. The level choice coincides with the launch of the updated versions of FCE (B2) and CAE (C1) in December 2008.

The post workshop tasks aimed at evaluating the effectiveness of the familiarisation activity and the familiarisation materials proposed by the abovementioned manual. Participants were therefore asked to revisit their earlier classification of descriptors into CEFR levels building on the knowledge gained and the discussion that had taken place at the workshop. Participants were also asked to provide action plans for cascading the knowledge gained.

AAAL 2008 Annual Conference, Washington DC

On March 30 Nick Saville (Cambridge ESOL) and Tim McNamara (University of Melbourne) led an invited symposium on *Issues of language acquisition and assessment as related to migration and citizenship* at this major event which took place in Washington DC from 29 March–1 April 2008. This symposium focused on language acquisition, learning and assessment as they relate to migration and citizenship. A range of theoretical and practical issues were presented including discussion of potentially positive impacts of language assessment and of negative impacts and of misuses of tests for these purposes.

The presenters incoluded the organisers together with Joe LoBianco (University of Melbourne), Elana Shohamy (Tel Aviv University), Piet Van Avermaet (Centre for Intercultural Education, University of Ghent), Anthony Kunnan (California State University, Los Angeles) and Tzahi Kanza (School of Education, Tel Aviv University). The discussants were Alan Davies (University of Edinburgh) and James Lantolf (The Pennsylvania State University).

In the introductory paper, Nick Saville presented on the current perspectives on migration and integration which point to a need to develop a coherent framework for addressing the assessment dimension when considering languages in society. This framework needs to be interdisciplinary and grounded in current theory and research on learning, teaching and the assessment of languages.

Next, Joe LoBianco spoke on *Denization*, *naturalization*, *education and freedom: Becoming a citizen*, in which he discussed citizenship as part of normative political philosophy, related to language expectations of four categories: the young (taught standard official language forms and standard literacy), the foreigner (classically linguistic socialisation is determined by occupational

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

position) and indigenous and enslaved minorities in whose cases marginalisation and compulsory illiteracy have been practised.

In the central part of the symposium, the five remaining presenters spoke under the heading *Language testing perspectives: What is the construct?* They focused on key issues from different parts of the world. They questioned how language testers can interact with other research strands in applied linguistics in considering the social and cultural dimensions with a view to identifying their own roles and distinctive voices. Ways of addressing the policy dimension with a view to dealing with language learning and assessment practices more effectively were discussed.

National Council on Measurement in Education 2008, New York

Ardeshir Geranpayeh (Cambridge ESOL) ran a joint security workshop on March 23, 2008. This was a pre-conference workshop for the National Council on Measurement in Education 2008 annual meeting in New York. The workshop, entitled *Test Security: Practices, Policies and Punishment,* was run jointly with James Impara and Jamie Mulkey from Caveon Test Security. Test security is a growing concern for learning institutions, credentialing organisations, and businesses. Each week, news stories with incidents of cheating, student coaching, teacher intervention, and even outright test theft are exposed. While there is an increase in these activities, new tools and methods are being developed to detect testing irregularities that are most likely caused by test fraud and theft.

Ardeshir Geranpayeh, Jamie Mulkey and James Impara at their NCME test security workshop.

The session took a case study approach to solving test security issues. Participants first gained an understanding of the impact of test theft on test takers and constituents. They were then given a primer on statistical analysis techniques used to detect answer copying and test administration irregularities. Using the results of statistical analysis techniques, participants used a case study to make decisions about applied policies and sanctions.

EAQUALS 2008 Conference, Paris

Hanan Khalifa and Evelina Galaczi led a workshop entitled *Exemplifying The CEFR: The creation and use of a DVD of students' oral language* at the EAQUALS 2008 Conference in Paris on the 24th–25th April. The conference theme was Adding Value in Language Education. EAQUALS is the European Association for Quality Language Services, a pan-European association of language training providers (see www.eaquals.org for more information). Cambridge ESOL is an Associate Member of EAQUALS.

Hanan and Evelina started the workshop by sharing with participants the rationale behind the project, how they selected oral samples, the instruments they used, and the procedures they carried out. The aim of the project was to provide typical oral performances illustrating CEFR levels A2 to C2 within a test-taking context. The selected performances are intended to be used as samples in standardisation training and ultimately in aiding a common understanding of the CEFR. They then attempted to replicate with participants what was done with the raters in the project in order to provide a practical illustration of the use of the DVD for training purposes. This included the participants' familiarisation with selected CEFR scales, and rating performances using CEFR scales. Participants were asked firstly to self-assess their own oral proficiency in a foreign language using CEFR scales and then to rate one of the oral performances from the DVD using both global and analytical rating scales, followed by a comparison and justification of the group's and individual's ratings. The workshop ended with a discussion of participants' reflections on the exercise especially on the application of the CEFR scales to the oral samples as well as sharing the findings from the study. It was intended that participants would take away both the tools and a methodology for replicating the exercise at a local level and the knowledge of what to consider when compiling samples of oral performances.

Twenty participants attended this event, some of whom are pictured below with the workshop presenters.

Hanan Khalifa (front row, second from left) and Evelina Galaczi (front row, fifth from left) at EAQUALS 2008 with workshop participants.

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

Research Notes offprints issues 1–32

Every issue of *Research Notes* is made available as a set of individual articles which are downloadable from our website. We now have over 200 offprints available on a wide range of topics, all written in an accessible style. These offprints are searchable by exam, skill or area and form a valuable free resource for our readers at www.cambridgeesol.org/rs_notes/offprints

We list below all of the offprints available to date. Note that issues 7 and onwards were themed issues.

Issue/Theme	Date	Title	Author/s
Issue 1	Mar 2000	EFL Research at UCLES	
		The EFL Local Item Banking System	Simon Beeston
		Developing Language Learning Questionnaires (LLQs)	Nick Saville
		Issues in Speaking Assessment Research	Lynda Taylor
		The UCLES/CUP Learner Corpus	Andrew Boyle, David Booth
Issue 2	Aug 2000	Stakeholders in language testing	Lynda Taylor
		Investigating the impact of international language examinations	Nick Saville
		The UCLES EFL item banking system	Simon Beeston
		Development of new item-based tests: The gapped sentences in the revised CPE Paper 3	David Booth, Nick Saville
		Background to the validation of the ALTE 'Can-do' project and the revised Common European Framework	Neil Jones
		Investigating the paired speaking test format	Lynda Taylor
		Using observation checklists to validate speaking-test tasks	Nick Saville
Issue 3	Nov 2000	Principles and practice in test development: the PETS Project in China	Lynda Taylor
		The use of Rasch Partial Credit Analysis in test development	Simon Beeston
		Developing observation checklists for speaking tests	Nick Saville, Barry O'Sullivan
		BULATS: A case study comparing computer based and paper-and-pencil tests	Neil Jones
		Approaches to rating scale revision	Lynda Taylor
		New-style statements of results	Neil Jones
Issue 4	Feb 2001	Reliability as one aspect of test quality	Neil Jones
		Test Development and Revision	Nick Saville
		Revising the IELTS Speaking Test	Lynda Taylor, Neil Jones
		Announcement of the winners of the IELTS MA Dissertation Award 2000	
Issue 5	Jul 2001	Revising the IELTS Speaking Test: developments in test format and task design	Lynda Taylor
		The ALTE Can Do Project and the role of measurement in constructing a proficiency framework	Neil Jones
		Towards a common scale to describe L2 writing performance	Roger Hawkey
		CB BULATS: Examining the reliability of a computer based test using test-retest method	Ardeshir Geranpayeh
Issue 6	Nov 2001	Issues in the assessment of second language writing	Stuart Shaw
		Using corpora in language testing	Fiona Ball
		Revising the IELTS Speaking Test: retraining IELTS examiners worldwide	Lynda Taylor
		The IELTS Impact Study: development and implementation	Roger Hawkey
		The paired speaking test format: recent studies	Lynda Taylor
		European language testing in a global context	Marianne Hirtzel
Issue 7 Testing young learners	Feb 2002	Developing English language tests for young learners	Lynda Taylor, Nick Saville
		Candidate performance in the Young Learners English Tests in 2000	Helen Marshall, Mike Gutteridge
		Research projects relating to YLE Speaking Tests	Fiona Ball, Juliet Wilson
		Striving for fairness – the ALTE Code of Practice and quality management systems	Nick Saville
		Investigating variability in a test of second language writing ability	Barry O'Sullivan
		Review of KET and PET Examinations	Nigel Pike, Liz Gallivan
		Report on the BAAL/CUP Seminar 'Young Language Learners: Towards a Research Agenda'	Fiona Ball
Issue 8 Testing English for business	May 2002	Some theoretical perspectives on testing language for business	Barry O'Sullivan
		Revising the Business English Certificates (BEC) speaking tests	David Booth
		Revising the BULATS Standard Test	Ed Hackett
		Developing wordlists for BEC	Fiona Ball
		The effect of training and standardisation on rater judgement and inter-rater reliability	Stuart Shaw
		Investigating gender differences in young learner performance	
		Investigating test conditions for listening and speaking	
		IELTS joint-funded research program: 1995–2001	

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

Issue/Theme	Date	Title	Author/s
Issue 9 **Certificates in English Language Skills (CELS)**	Aug 2002	Plurilingualism, partial competence and the CELS suite	Lynda Taylor
		Background to CELS: the communicative construct and the precursor exams	Roger Hawkey
		The test development process for CELS	Nick Saville
		CELS Writing: test development and validation activity	Stuart Shaw, Sharon Jordan
		CELS Speaking: test development and validation activity	Lynda Taylor, Stuart Shaw
		IELTS Writing: revising assessment criteria and scales (Phase 1)	Stuart Shaw
		Investigating the CPE word formation cloze task	
		Reviewing the retraining of BEC Oral Examiners	
Issue 10 **Exam reviews: CPE, KET, PET, IELTS, YLE**	Nov 2002	Innovation and continuity: CPE – past and present	Cyril Weir
		Redeveloping Part 1 of the CPE Listening paper	Rod Boroughs
		Update on changes to the KET/PET Writing papers from 2004	Liz Gallivan
		IELTS Writing: revising assessment criteria and scales (Phase 2)	Stuart Shaw
		Linking YLE levels into a single framework	Neil Jones
		Investigating the YLE story-telling task	Fiona Ball
		Assessing learners' English: but whose/which English(es)?	Lynda Taylor
		Exploring issues in the assessment of pen-and-paper/computer-based IELTS Writing	
		Lexicom@ITRI: a Lexicography Course	
		Monitoring oral examiner performance in FCE	
		Monitoring IELTS test performance in 2001	
		Monitoring speaking test materials for Young Learners Tests	
Issue 11 **Testing candidates with special needs**	Feb 2003	Responding to diversity: providing tests for language learners with disabilities	Lynda Taylor, Mike Gutteridge
		Producing Modified Versions of Cambridge ESOL Examinations	Ruth Shuter
		Legibility and the rating of second language writing: the effect on examiners when assessing handwritten and word-processed scripts	Stuart Shaw
		Task difficulty in the assessment of writing: Comparing performance across three levels of CELS	Neil Jones, Stuart Shaw
		Insights into the FCE Speaking Test	Yang Lu
		Converting an Observation Checklist for use with the IELTS Speaking Test	Lindsay Brooks
Issue 12 **Technology in language testing**	May 2003	The Role of Technology in Language Testing	Neil Jones
		Electronic Script Management: towards on-screen assessment of scanned paper scripts	Stuart Shaw
		A quick review of the English Quick Placement Test	Ardeshir Geranpayeh
		Recent Developments in Learner Corpora	Fiona Barker
		Assistive Technology for Candidates with Special Needs	Mike Gutteridge
		Feedback on CPE re-training	Chris Hubbard
Issue 13 **Testing speaking**	Aug 2003	The Cambridge approach to speaking assessment	Lynda Taylor
		What constitutes a basic spoken vocabulary?	Michael McCarthy, Ronald Carter
		The development of a set of assessment criteria for Speaking Tests	Angela ffrench
		CELS Speaking Assessment: towards an understanding or oral examiner and test-taker behaviour	Stuart Shaw
		Evaluating the success of the revised BEC (Business English Certificate) Speaking Tests	David Booth
Issue 14 **Teaching Awards**	Nov 2003	Cambridge ESOL Teaching Awards: current perspectives, future trends	Monica Poulter
		The Distance DELTA	David Albery
		DELTA by Distance Learning	Dave Russell
		Diaries, theory, practice and assessment: the teacher educator as reflective practitioner	Craig Thaine
		Language Awareness and Assessment	Pauline Rea-Dickins
		In-service language teaching in Brazil using ICELT	Lizika Goldchleger
		Teacher Support	Jill Grimshaw
		Interaction in a paired speaking test: the case of the First Certificate in English	Evelina Galaczi
Issue 15 **Testing Language for Specific Purposes**	Feb 2004	Issues of test comparability	Lynda Taylor
		Analysing domain-specific lexical categories: evidence from the BEC written corpus	David Horner, Peter Strutt
		IELTS Writing: revising assessment criteria and scales (concluding Phase 2)	Stuart Shaw
		An IELTS Impact Study: implementation and some early findings	Roger Hawkey
		The YLE Review: findings from a stakeholder survey	Trish Burrow, Juliet Wilson
		Creating a virtual community of assessment practice: towards 'on-line' examiner reliability	Stuart Shaw
		Reliability in First Certificate in English objective papers	Ardeshir Geranpayeh
		Announcement of the winner of the IELTS Masters Award 2003	
Issue 16 **Testing writing**	May 2004	Second language writing assessment: Cambridge ESOL's ongoing research agenda	Lynda Taylor
		IELTS Writing: revising assessment criteria and scales (Phase 3)	Stuart Shaw
		Exploring the relationship between YLE Starters and Movers and Breakthrough level	Trish Burrow
		Making the grade: score gains on the IELTS Writing test	Tony Green
		Question uptake in the Certificate in Advanced English Writing Paper	Fiona Barker, Cris Betts

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

Issue/Theme	Date	Title	Author/s
Issue 17 **Language testing in Europe**	Aug 2004	A Common Solution to a Common European Challenge: The work of ALTE	Barbara Stevens
		Test Equivalence and Construct Compatibility across Languages	Peter Hardcastle
		Development of an Electronic European Language Portfolio	Simon Fenn
		Automated Writing Assessment: a review of four conceptual models	Stuart Shaw
Issue 18 **IELTS**	Nov 2004	IELTS, Cambridge ESOL examinations and the Common European Framework	Lynda Taylor
		Computer-based IELTS and paper-based versions of IELTS	Tony Green, Louise Maycock
		IELTS Impact: a study on the accessibility of IELTS GT Modules to 16–17 year old candidates	Jan Smith
		IELTS Writing: revising assessment criteria and scales (Phase 4)	Graeme Bridges, Stuart Shaw
		Set Texts in CPE Writing	Diana Fried-Booth
		IELTS – some frequently asked questions	
		IELTS test performance data 2003	Andrew Blackhurst
Issue 19 **Development of assessment products: TKT, Asset**	Feb 2005	Rising to the Challenge of Asset Languages	Neil Jones, Karen Ashton
		Opening a new door for teachers of English: Cambridge ESOL Teaching Knowledge Test	Mick Ashton, Hanan Khalifa
		Staying in Touch: tracking the career paths of CELTA graduates	Tony Green
		Cambridge ESOL and the NRDC ESOL Effective Practice Project	James Simpson
		Raising the Languages Ladder: constructing a new framework for accrediting foreign language skills	Neil Jones
		The Common Scale for Writing Project: implications for the comparison of IELTS band scores and Main Suite exam levels	Roger Hawkey, Stuart Shaw
Issue 20 **Impact on stakeholders**	May 2005	Washback and impact: the view from Cambridge ESOL	Lynda Taylor
		The effects on performance of computer familiarity and attitudes towards CB IELTS	Louise Maycock, Tony Green
		Skills for Life writing mark scheme trial: validating the rating scale for Entry Levels 1, 2 and 3	Stuart Shaw, Evelina Galaczi
		Applying lexical statistics to the IELTS speaking test	John Read
		Upper Main Suite speaking assessment: towards an understanding of assessment criteria and oral examiner behaviour	Evelina Galaczi
		The CPE Textbook Washback Study	Roger Hawkey
		IELTS joint-funded research program: 2002–2004	
Issue 21 **Developing materials for language tests**	Aug 2005	Using qualitative research methods in test development and validation	Lynda Taylor
		Quality Assurance and Quality Control: Reviewing and pretesting examination material at Cambridge ESOL	Tony Green, David Jay
		Are test taker characteristics accounted for in Main Suite Reading papers?	Hanan Khalifa
		Establishing the validity of Cambridge ESOL Writing Tests: towards the implementation of a socio-cognitive model for test validation	Cyril Weir, Stuart Shaw
		Listening, Reading and Writing on computer-based and paper-based versions of IELTS	Andrew Blackhurst
Issue 22 **Ethics in testing**	Nov 2005	Setting and monitoring professional standards: a QMS approach	Nick Saville
		Ethical issues in the testing of young learners	Juliet Wilson
		An overview of computer-based testing	Paul Seddon
		The development of a computer-based version of PET	Ed Hackett
		Evaluating the impact of word processed text on writing quality and rater behaviour	Stuart Shaw
Issue 23 **Technology in language testing**	Feb 2006	Assessment systems: conceptual, human, technological	Neil Jones
		The Cambridge ESOL Item Banking System	Helen Marshall
		ESOL Professional Support Network Extranet	Clare Mitchell Crow, Chris Hubbard
		IELTS Writing: revising assessment criteria and scales (Phase 5)	Peter Falvey, Stuart Shaw
		IELTS test performance data 2004	
		IELTS award news 2005	
		ESOL Special Circumstances 2004: A review of Upper Main Suite provision	Mike Gutteridge
Issue 24 **Frameworks in assessment**	May 2006	Cambridge ESOL exams and the Common European Framework of Reference (CEFR)	Lynda Taylor, Neil Jones
		Placing the International Legal English Certificate on the CEFR	David Thighe
		Linking learners to the CEFR for Asset Languages	Tamsin Walker
		Can Do self-assessment: investigating cross-language comparability in reading	Karen Ashton
		Assessment processes in Speaking tests: a pilot verbal protocol study	Chris Hubbard, Susan Gilbert
		IELTS Writing: revising assessment criteria and scales (Conclusion)	Stuart Shaw
		TKT – a year on	Nadežda Novaković
Issue 25 **Testing Language for Specific Purposes**	Aug 2006	Language testing for migration and citizenship	Nick Saville
		Issues with developing a test in LSP: the International Certificate in Financial English	Kate Ingham, David Thighe
		Using the global legal community in the development of ILEC	David Corkill, Martin Robinson
		Developing the Cambridge ESOL Teacher Portfolio	Clare Mitchell Crow, Clare Harrison
		Profile of Skills for Life candidature	Nadežda Novaković
		The impact of proficiency-level on conversational styles in paired speaking tests	Fumiyo Nakatsuhara

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

Issue/Theme	Date	Title	Author/s
Issue 26 Corpora and language assessment	Nov 2006	Corpora and language assessment: trends and prospects	Fiona Barker
		Developing a classroom video database for test washback research	Roger Hawkey, Sue Thompson
		Defining the constructs underpinning the Main Suite Writing Tests: a socio-cognitive perspective	Cyril Weir, Stuart Shaw
		A worldwide survey of examiners' views and experience of the revised IELTS Speaking test	Annie Brown, Lynda Taylor
		The effect of editing on language used in FCE reading texts: a case study	Glyn Hughes
Issue 27 Testing English for business	Feb 2007	Cambridge ESOL and tests of English for Specific Purposes	David Thighe
		Publishing vocabulary lists for BEC Preliminary, PET and KET examinations	Jason Street, Kate Ingham
		Using simulation to inform item bank construction for the BULATS computer adaptive test	Louise Maycock
		The comparability of computer-based and paper-based tests: goals, approaches, and a review of research	Neil Jones, Louise Maycock
		Modelling facets of the assessment of Writing within an ESM environment	Stuart Shaw
		Broadening the cultural context of examination materials	Steve Murray
		IELTS Masters Award 2006	
Issue 28 Testing young learners	May 2007	Reviewing the Cambridge Young Learners English (YLE) tests	Juliet Wilson
		The marking of spelling for the revised YLE tests from January 2007	Helen Spillett
		Cambridge ESOL YLE tests and children's first steps in reading and writing in English	Shelagh Rixon
		Linking language assessments for younger learners across proficiency levels (Phase 1)	Fiona Barker, Stuart Shaw
		IELTS Joint-funded Research Program: Rounds 1–12	
Issue 29 Teaching Awards	Aug 2007	Cambridge ESOL teacher training and development – future directions	Monica Poulter
		The DELTA Revision Project – progress update	Ron Zeronis
		DELTA reliability: estimating and reporting examiner performance indices for the written examination component	Stuart Shaw
		What difference does DELTA make?	Simon Phipps
		Setting international standards for teaching	Monica Poulter
		Communities of practice and teacher education: the contribution of the CELTA trainer training programme	Jo-Ann Delaney
		ICELT and PEP Ukraine: evaluation of a reflective ESP teacher development programme	David Watkins
		Teaching Knowledge Test update – adoptions and courses	Clare Harrison
Issue 30 Exam reviews: FCE, CAE	Nov 2007	The 2004–2008 FCE and CAE Review Project: historical context and perennial themes	Roger Hawkey
		Using Structural Equation Modelling to facilitate the revision of high stakes testing: the case of CAE	Ardeshir Geranpayeh
		Introducing short themed texts into the CAE Reading paper	Helen Coward
		Establishing the impact of reduced input and output length in FCE and CAE Writing	Margaret Cooze, Stuart Shaw
		Reviewing the CAE Listening test	Steve Murray
		Reviewing Part 1 of the FCE Listening test	Diana Fried-Booth
		Reviewing the FCE and CAE Speaking tests	Clare Harrison
		Developing revised assessment scales for Main Suite and BEC Speaking tests	Evelina Galaczi, Angela ffrench
		Overview of FCE and CAE Review Project research activity	Fiona Barker, Steve Murray
Issue 31 Testing reading	Feb 2008	A cognitive processing approach towards defining reading comprehension	Cyril Weir, Hanan Khalifa
		Applying a cognitive processing model to Main Suite Reading papers	Cyril Weir, Hanan Khalifa
		A corpus-informed study of specificity in Financial English: the case of ICFE Reading	Angela Wright
		Exploring lexical differences in General English Reading papers	Fiona Barker
		Text organisation features in an FCE Reading gapped sentence task	Glyn Hughes
		IELTS award news	
Issue 32 Testing listening	May 2008	Examining Listening: developments and issues in assessing second language listening	Ardeshir Geranpayeh, Lynda Taylor
		The Cambridge ESOL approach to item writer training: the case of ICFE Listening	Kate Ingham
		Vocabulary use in the FCE Listening test	Dittany Rose
		Using DIF to explore item difficulty in CAE Listening	Ardeshir Geranpayeh
		Adapting listening tests for on-screen use	Ed Hackett
		Research Notes offprints issues 1-32	

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.

Cambridge to host 34th IAEA Annual Conference

Leading education experts and exam providers from across the world will exchange the latest research, ideas and experiences of different countries' education systems at a major conference in Cambridge.

The 34th International Association for Educational Assessment (IAEA) Annual Conference is being hosted by Cambridge ESOL's parent body, Cambridge Assessment, as part of the celebrations for its 150th anniversary. The event will take place at Cambridge University's Robinson College from 7–12 September 2008. The call for papers has now closed, however registration is open until 11 July.

Internationally recognised education thinkers Professor Robert J Mislevy, of the University of Maryland, and Professor Dylan Wiliam, of the Institute of Education, London, are the keynote speakers. They will present challenging views on the way assessment has changed and the directions it may take in the future. The main conference theme, *Re-interpreting Assessment: Society, Measurement and Meaning*, will encourage debate on technical measurement issues and how results from assessments are used in the wider world.

Group Chief Executive of Cambridge Assessment, Simon Lebus, said: 'Effective assessment enriches lives, broadens horizons and shapes futures, and shows us when education works. The skills people require to succeed keep changing, and assessment must evolve to keep pace. We want to ensure individuals continue to progress – assessment must inform the learning process. We very much look forward to welcoming participants to Cambridge.'

The sub themes include *using technology in assessment, standards setting, multiculturalism and assessment* and *equity issues in assessment*. Participants will be able to consider the main issues, challenges and developments in the field of assessment today.

The main conference sessions will be held at Robinson College, Cambridge University's newest college and home of the Prince of Wales' education summer school in 2006. The college is an architecturally striking, highly functional building set in several acres of attractive wooded gardens close to the city centre. Social functions will include a welcome reception at Robinson College and a gala dinner in the magnificent Great Hall at King's College, one of Cambridge University's oldest colleges, founded in 1441 by Henry VI.

For further information and to register your place, please visit www.iaea2008.cambridgeassessment.org.uk

IELTS Joint-Funded Research Programme 2008–9

The IELTS partners are once again making available grant funding for IELTS-related research projects to be conducted during 2009. The total value of funds available is £120,000 (AUS$225,000) with a maximum of £15,000/AUD $36,000 per selected project.[1]

Institutions/individuals are invited to submit a written application (maximum of 10 pages) in accordance with the format and content requirements provided on the IELTS website. Full details of the IELTS research programme, current areas of research interest and funding application guidelines are available at: www.ielts.org/teachers_and_researchers/grants_and_awards/

All applications received will be treated on a confidential basis. The decision of the review committee will be final.

Closing date for receipt of completed application forms and research proposals is 30 June 2008.

Signed applications should be submitted by both email and post to the following contacts:

British Council
Paul Wade
IELTS Marketing Officer
British Council
English and Exams
10 Spring Gardens
London SW1A 2BN
Tel. +44 (0)20 7389 3140
Fax. +44 (0)20 7389 4140
Email paul.wade@britishcouncil.org

IDP Australia
Marcia Caswell
Regional Manager, IELTS
IDP Education Pty. Ltd
GPO Box 2006
ACT 2601
Australia
Tel. +61 26285 8372
Fax. +61 26285 3233
Email marcia.caswell@idp.com

[1] This upper limit may be reviewed in exceptional circumstances.

©UCLES 2008 – The contents of this publication may not be reproduced without the written permission of the copyright holder.