ResearchNotes

Contents

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

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

Editorial Notes

Welcome to issue 26 of *Research Notes*, our quarterly publication reporting on matters relating to research, test development and validation within Cambridge ESOL.

The theme of this issue is corpora and language assessment, which is an increasingly important aspect of language testing and related areas such as teaching and publishing, as well as being more widely used in diverse fields within linguistics and education. In this issue we provide an overview of the use of corpora in testing to date and describe our current involvement in the development of corpus resources whilst also considering how these and other technological developments (such as Electronic Script Management, ESM) inform our understanding of the constructs underlying language tests. We report on some successful corpus-informed test development activities and describe a number of case-studies which have explored Reading texts and Spoken data using diverse corpus approaches.

In the opening article Fiona Barker outlines the growth in the existence and use of corpora for language assessment and describes a range of current corpus-related activities before looking to future applications of this field for language testers, both within Cambridge ESOL and more widely. Next, Roger Hawkey, Sue Thompson and Richard Turner describe the development of a video database of classroom data from three impact studies which will aid research into test washback for a range of different exams and contexts. This database is a form of multimodal corpus containing video clips, metadata and subtitling.

In the following article Cyril Weir and Stuart Shaw summarise the constructs underpinning the Main Suite Writing tests, drawing in part on corpus evidence. A clear construct definition is vital for understanding and validating language tests and this article describes the application of a socio-cognitive validity framework to the Cambridge Writing examinations.

Annie Brown and Lynda Taylor report on a project commissioned by Cambridge ESOL to survey examiners' views and experience of the IELTS Speaking Test following its revision in 2001. This survey explored both the Speaking Test's format and tasks and how raters assessed candidates taking the test.

The following article describes corpus-based research undertaken for a Masters level course. Glyn Hughes compares texts used in the Reading component of FCE and the British National Corpus (BNC) to ascertain what impact edited Reading texts have on candidates, basing his article on a case study of the word *people*.

We end this issue with reports from conferences and other events attended by Cambridge ESOL staff in the past year, several of which included papers related to corpus-informed research. We look forward to the coming year, including a new look for *Research Notes*, and would like to wish all of our readers a happy, healthy and prosperous 2007.

Corpora and language assessment: trends and prospects

FIONA BARKER, RESEARCH AND VALIDATION GROUP

Introduction

This article describes how Cambridge ESOL is developing and using corpora and looks to future applications of these resources within language assessment more widely.[1] In the 1970s and 80s corpora provided applied linguists with valuable insights into the grammatical, lexical and discourse features of native speaker (NS) English, although they did not impact significantly on the language assessment community. By the 1990s, however, corpora such as the COBUILD Bank of English and the British National Corpus (BNC) were shaping developments in applied linguistics and language pedagogy through the publication of reference works and the use of corpora in teaching.

Since the early 1990s Cambridge ESOL has been developing collections of written, and more recently, spoken, learner output which serve as archives as well as having numerous research and operational uses (see Barker 2004). Our best-known corpus is the Cambridge Learner Corpus (CLC), a 25 million word collection of candidates' written responses. The CLC is amongst the largest learner corpora worldwide and was developed to inform test development and publishing activities for Cambridge ESOL and Cambridge University Press. Cambridge ESOL has direct control over the CLC's design and contents; we ensure, for example, that sampled scripts cover a range of proficiency levels and we also target specific first language groups (L1s) for each exam. Every year we collect a representative sample of scripts for the CLC, this selection being based on the sometimes conflicting demands of maintaining a representative spread of L1s and exams and building up similar amounts of data across different proficiency levels and language domains (i.e. general, business, academic). Scripts are keyed in to form a computer-readable version and are accompanied by candidate information and score data so that users can filter searches by variables such as age, L1 or exam grade. The CLC is searchable through proprietary software either by error types (thanks to the unique error-tagging system, see Nicholls 2003), or lexically through a concordancer or collocation search. Cambridge ESOL adds new exams and functionality to the CLC based on our research and operational requirements together with the growth in certain markets and domain-specific testing. Today the CLC includes 16 different tests; future additions may include the ESOL Skills for Life exams (SfL) and domain-specific tests of legal and financial English.[2]

Using corpora at Cambridge ESOL

The CLC has provided data for a range of research and operational activities, described in Barker (2004). Operational research has included updating item writer and syllabus word lists for various exams (Ball 2002a); informing the Common Scale for Writing (Hawkey and Barker 2004); and analysing candidates' business lexis using frequency word lists (Horner and Strutt 2004). Longer-term research has compared candidates' written and spoken vocabulary with existing item writer word lists, and investigated the influence of varieties of English on candidates' written vocabulary (Taylor and Barker 2006). The CLC is currently being used to develop a specification of vocabulary, structures and other features of writing at the six proficiency levels (A1–C2) of the CEFR (Council of Europe 2001). These findings will feed into a much larger, collaborative project to produce a comprehensive reference level description called English Profile (www.EnglishProfile.org). Cambridge ESOL also develops smaller collections of learner or NS writing for specific research projects or operational procedures.

Whilst written corpora have proved beneficial, Cambridge ESOL has also collected spoken learner performances, although this has not been straightforward due to the challenges of recording and annotating speech. We have, however, collected and catalogued thousands of speaking tests from our existing archive and special recording sessions – around 5% of which have been transcribed or digitised – and we are investigating the best way to store these longer-term; one option is to use a database approach like that used by Hawkey et al for the Impact Video Database (see p5). The spoken corpus is not yet searchable like the CLC, although it has provided data for a number of external researchers and projects funded by the IELTS Joint-funded Research Program (see www.ielts.org).[3] Examples of research based on the spoken corpus include:

- children's language use in the Cambridge YLE Tests (Ball 2002b, Ball and Wilson 2002)
- interaction in FCE Speaking Tests of various L1 groups (Galaczi 2003)
- applying lexical statistics to candidate output in revised IELTS speaking tests (Read 2005).

Alongside their research uses, corpora are also used operationally at Cambridge ESOL. For example, some of our question paper writers (item writers) routinely use the concordance function of a general NS corpus such as the BNC – often through an online sampler – or a web search engine for various activities including:

writing items:

- to establish authentic contexts for a target item (content and language)

[1] This article contains extracts from a chapter in preparation for Springer's *Encyclopedia of Language and Education* (Taylor and Barker forthcoming) and also draws on contributions from Cambridge ESOL colleagues (thanks to Trish Burrow, Diana Fried-Booth, John Picton and Judith Wilson).

[2] See www.cambridge.org/elt/corpus/learner_corpus.htm for details of the size, content and uses of the CLC.

[3] Researchers can find details of how to submit a request for access to data or materials at www.CambridgeESOL.org/research

- to generate language which reflects natural authentic usage
- to find the most common form for a target item (e.g. singular or plural noun)
- to check whether their intuitions about an aspect of language to be tested are correct.

editing or proof-reading items:

- to check the frequency/authenticity of a collocation, idiom etc., or the appropriacy of a given context
- to establish whether an item should be tested, given its frequency or authenticity
- to check word senses or 'shades of meaning' which may not always appear in a dictionary
- to check whether distractors could be possible keys in certain varieties of English.

An item writer notes that 'a turn of phrase which seems quite acceptable and is grammatically accurate can turn out to have limited contemporary usage' and suggests that 50 citations for a selected word 'supply a sufficient number of items in context which can usually confirm whether the item will provide a fair target for testing purposes'. Cambridge ESOL subject officers also have access to the CLC (including derived frequency word lists) and NS corpora for checking question papers. A subject officer maintains that '[corpora] have proved to be invaluable when proofreading or checking items that are clearly designed to test candidates' knowledge of collocations…gaining a good number of 'hits' for the collocation gives one an assurance of the 'validity' of the collocation suggested by the item and the key'. Clearly, corpora are being used in a range of ways within Cambridge ESOL but we should also consider their benefits and risks and potential uses within language assessment more widely.

Corpus-informed language testing: benefits and risks

Charles Alderson (1996) signalled a potential role for corpora in language assessment in the mid-1990s and with an increasing range of corpora available, particularly learner corpora (see Granger 2004 and Pravec 2002), language testers began actively exploring their application. Cambridge ESOL, for example, drew on the BNC, COBUILD and CLC to inform test revision (Weir and Milanovic 2003) and to devise new test formats (Hargreaves 2000). In the USA, the *TOEFL 2000 Spoken and Written Academic Language Corpus* (T2K-SWAL) was developed to investigate university-level language skills and provides an empirically-grounded alternative to the intuitions of TOEFL test constructors and writers (Biber et al 2004).

There are a number of advantages of using corpora within language testing. Corpora of language test content (input) and of test taker performance (output) provide language testers with archives that enable them to address issues such as comparability across test forms, rater training and standardisation, standard-setting and maintenance over time, and investigation of test bias across test taker populations. Small-scale, specialised corpora are also valuable as these provide insights into task- or domain-specific issues. Large online corpora mean that anyone can access them with relative ease and user-friendly interfaces afford either a 'quick and dirty' look into relevant data or the basis for deeper investigations. There are also many instructional publications and software available to aid corpus investigation. Corpus analyses allow detailed comparison across tests at different proficiency levels or for different domains in terms of lexical, structural, and functional content; they can also assist in placing tests within a larger framework of reference such as the CEFR and developing performance descriptions in the form of Can-Do statements (Council of Europe 2001).

Using corpora to inform language assessment also brings a number of potential risks. Whilst different corpora provide multiple points of reference, it is important to note that every corpus is designed along differing parameters and purposes which constrain what it can be used for. The CLC, for example, was designed for general and specific test development/validation projects across many different tests whilst T2K-SWAL was designed for the revision of one specific academic test. The representativeness and relevance of a corpus, including its size and age, therefore need to be considered when using it to inform decision-making (Alderson 1996). Corpus evidence is rarely conclusive, hence why language testers tend to triangulate corpus-evidence with other forms of enquiry and analysis (Hawkey and Barker 2004). Nevertheless, corpus evidence can indicate where on a cline (whether of acceptability, frequency etc.) a specific language feature lies and provides real examples in use. Interpreting corpus data requires the same care as the interpretation of statistical analyses; this can be challenging where the corpus data are strongly influenced by a task effect, which is true for any corpus of test taker performance.

In 2003 a symposium at the *Language Testing Research Colloquium* in Reading (Taylor et al 2003) considered the ways in which corpora were becoming increasingly useful to the language testing community. Some of the likely future applications of corpora are considered below.

Future applications of corpora for language testing

Alderson (ibid) proposed a range of corpus applications in language assessment including test writing, construction, scoring and score reporting, although he cautioned against being seduced by the 'cleverness' of new technology and encouraged language testers to keep in mind fundamental theoretical considerations (e.g. construct definition, validity, reliability) alongside empirical findings. The key areas for future developments are discussed in Taylor and Barker (forthcoming) and are summarised below.

Test scoring and rating

An area of current growth is the automated evaluation of essays and short answer responses in content-based tests. In 1999 ETS introduced software for scoring essays holistically and Burstein et al (2001) reported on the automated scoring of short answer responses. The automated scoring of spoken performance is a more complex endeavour but rapid developments in computer-based speech technology are bringing this operational reality closer.

Use of new technologies

The spread of computer-based testing worldwide (e.g. computer-based IELTS) should permit easier corpus-building as test takers' written responses could feed straight into a corpus without the need for labour-intensive keying. Rapid advances in audio and speech recognition technology will make it easier to store, search and analyse speaking test data, including automatic transcription and tagging. Another application of corpus technologies to assessment concerns the development of software for detecting cheating and malpractice. The issues surrounding plagiarism are currently receiving increasing attention in the literature, and the use of plagiarism detection software is growing in academic and other high-stakes assessment contexts.

New types of corpora

Those corpora of increasing interest to language testers will be those which expand and branch out into new directions, namely field-specific reference corpora (e.g. for law or accountancy) and age-specific corpora as the assessment of younger language learners increases. Other corpora focusing on international language varieties could inform decisions about the level of linguistic variation to be included in language tests (Taylor 2006).

Conclusion

Corpora are increasingly recognised as an important research and operational tool for language assessment and this interest is reflected in increased reporting at conferences and in the literature. The current use of corpora varies according to a language testing institution's ethos and the types of language tests they offer, also according to the needs, knowledge and expertise of their staff involved in test production and research. The applications of corpora described here are revealing more detailed information about language of interest to language testers. However, more sophisticated methods of data analysis need to be found to support future language testing endeavours.

There is clearly a range of initiatives currently taking place at the interface of corpus linguistics and language assessment and this article has sought to demonstrate that corpora are being applied to the evaluation of language proficiency and that this enterprise has a promising future. At Cambridge ESOL we continue to develop our corpus resources and are investigating new ways to utilise these, to keep in line with our range of language tests and teaching awards and the demands we wish to make of these unique resources for future language testing provision.

References and further reading

Alderson, J C (1996) Do corpora have a role in language assessment? in Thomas, J A and Short, M H (Eds) *Using Corpora for Language Research*, London: Longman.

Ball, F (2002a) Developing word lists for BEC, *Research Notes* 8, 10–13.

—(2002b) Investigating the YLE story-telling task, *Research Notes* 10, 16–18.

Ball, F and Wilson, J (2002) Research Projects relating to YLE Speaking Tests, *Research Notes* 7, 8–10.

Barker, F (2004) Using Corpora in Language Testing: Research and validation of language tests, *Modern English Teacher*, 13/2, 63–67.

Biber, D, Conrad, S, Reppen, R, Byrd, P, Helt, M, Clark, V, Cortes, V, Csomay, E and Urzua, A (2004) *Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus*, report Number: RM–04–03, Educational Testing Service, Princeton, NJ.

Burstein, J, Leacock, C and Swartz, R (2001) *Automated Evaluation of Essays and Short Answers*, paper presented at 5th International Computer Assisted Assessment Conference, Loughborough University, UK. www.caaconference.com/pastConferences/2001/proceedings/a.pdf

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

Galaczi, E (2003) Interaction in a paired speaking test: the case of the First Certificate in English, *Research Notes* 14, 19–23.

Granger, S (2004) Computer learner corpus research: current status and future prospects, in Connor, U and Upton, T A (Eds) *Applied Corpus Linguistics: A Multidimensional Perspective*, Rodopi: Amsterdam and Atlanta.

Hargreaves, P (2000) How Important Is Collocation in Testing the Learner's Language Proficiency?, in. Lewis, M (Ed) *Teaching Collocation – further developments in the Lexical Approach*, Language Teaching Publications: Hove.

Hawkey, R and Barker, F (2004) Developing a common scale for the assessment of writing, *Assessing Writing*, 9, 122–159.

Horner, D and Strutt, P (2004) Analysing domain-specific lexical categories: evidence from the BEC written corpus, *Research Notes* 15, 6–8.

Nicholls, D (2003) The Cambridge Learner Corpus - error coding and analysis for lexicography and ELT, in Archer, D, Rayson, P, Wilson, A and McEnery, T (Eds), *Proceedings of the Corpus Linguistics 2003 Conference*, UCREL technical paper number 16, UCREL, Lancaster University.

Pravec, N A (2002) Survey of learner corpora, *ICAME Journal* 26, 81–114.

Read, J (2005) Applying lexical statistics to the IELTS speaking test, *Research Notes* 20, 12–16.

Taylor, L and Barker, F (2006) Using corpus studies to explore the impact of regional varieties of English on learners' written performance, paper presented at BAAL 2006, Cork.

—(forthcoming), Using corpora for language assessment, in Hornberger, N H (Ed) (2 ed), Encyclopedia of Language and Education, Springer.

Taylor, L, Thompson, P, McCarthy, M and Barker, F (2003) *Exploring the relationship between language corpora and language testing*, Symposium at 25th Language Testing Research Colloquium, Reading. 22–25 July 2003.

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

Developing a classroom video database for test washback research

ROGER HAWKEY, CAMBRIDGE ESOL CONSULTANT
SUE THOMPSON, CAMBRIDGE ESOL VIDEO UNIT
RICHARD TURNER, RESEARCH AND VALIDATION GROUP

Introduction

Three recent Cambridge ESOL impact studies have used video-recorded classroom observation data as a key part of their evidence. The studies were of the International English Language Testing System (IELTS), of the *Progetto Lingue 2000* foreign language reform project in Italy, and of preparation courses for internal and external English language exams at a language centre in Florence. For language test providers such as Cambridge ESOL, studies of test impact play an integral role in systems for the continuous validation of its exams, in particular their *consequential validity* (Messick 1989). How do high-stakes tests, for example, wash back on the teaching, learning, materials and methods of courses preparing the test takers? What impact do these tests have on users such as language schools, ministries and receiving institutions?

This article reports on a project to develop a database of classroom videos to help test washback research.

Impact and washback

We accept here definitions (e.g. Bachman and Palmer 1996, Bailey 1996, Hamp-Lyons 2000) suggesting that *washback* occurs at the 'micro' level, that is affecting participants such as learners and teachers, while *impact* is a more 'macro' concept, involving stakeholders beyond the immediate learning/teaching context. But neither washback nor impact normally occur in linear cause and effect patterns. All three projects thus sought to take account of the complexity of relationships between language tests and the teaching and learning associated with them. Alderson and Hamp-Lyons (1996), Hawkey (2006), Spratt (2005) and Watanabe (1996) discuss significant differences in the ways in which, for example, different teachers handle classes preparing students for the same exam. It is clear that the more analysable and re-visitable data we have, 'enabling reflection on performance or behaviour' (Canning-Wilson 2000) the better we will be able to investigate the complex variables involved in test washback. Among such data, test-related classroom observation evidence, video-recorded and data-based, can play a key role in clarifying complicated washback issues.

The need, in washback studies, for classroom observation data 'incorporating first-hand evidence of classroom events' (Andrews 2004:50) is often emphasised in the literature (e.g. Alderson and Wall 1993, Green 2003, Hawkey 2006, Turner 2001), though Spratt's recent survey (ibid) of eleven empirical studies on washback suggests that only five of these used classroom observation. The video-recording of classroom activities is even more useful for researchers if there is access to the data in a systematically organised, labelled and retrievable computerised form. This article describes the development into an expandable and adaptable research database of the video-recorded classroom data from our three impact studies.

The use of video

From our survey of the literature on educational video use, the project team noted three major trends: a significant, though, it seems, still insufficient, increase, in the use of video in educational research (Brophy 2003, Goldman et al 2004, Hollingsworth 2005); a focus on the strengths and pitfalls of video use in learning and teaching across the curriculum and thirdly, in teacher education (Abell and Cennamo 2003, Benjamin, Brewer and Hebl 2000, Brophy ibid, Constantinou and Papadouris 2004, Cornfield, Campbell and McCammon 2001, Fishman 2003, Lowery 2005, Russow 2003, Seago 2003). In the ELT field Canning-Wilson (ibid), Dudeney (2000), Migliacci (2002), Rhodes and Pufahl (2003) and Stempleski and Arcario (1992) discuss the many and sophisticated ways of using video already being practised. The use of video in language testing is also a matter of current discussion. Jones (2001:3–4) refers to 'simulated oral proficiency interviews, where a candidate interacts with a recorded interlocutor and his performance is recorded for later assessment'. Saidatul (2006) illustrates video and CD-ROM technology to apply this principle to a university oral performance test. Recent articles on Cambridge ESOL's computer-based testing initiatives have included Blackhurst (2005), Green and Maycock (2004), Hackett (2005), Maycock and Green (2005) and Shaw (2005).

Video databases

Video databases are increasingly widely used for research purposes, often in the education sector; their aim is to make many hours of video-recording available through an organisational system which can tell users 'what is available that they would not have known to look for at precisely the moment when they might become interested in that information' (Institute for Learning Science 2006). Our inquiry into the relevant *video database* literature uncovered proposed solutions to problems such as video clip division into segments, and indexing the segments and representing the indices to offer the user the most convenient method of browsing and retrieval (e.g. Elmasri 2004, Furth and Marques 2003). These were the aspects of video databasing that were most relevant to the development of our own video database. Plowman (1999:6) notes an additional potential advantage of research video databases, namely that they may enable us to 'share and discuss emergent findings with other researchers and practitioners', or 'review interesting or problematic sections [with other researchers]… who can test the validity of our findings by analysing the same video material'. Some databases thus combine their systematicity and retrievability with the global accessibility of the World Wide Web, the ultimate in data-sharing. The constraints on such data sharing in our case are discussed below.

At its most basic, a video database stores video-related data, then either also stores a link to the video, or embeds the video object within the database. Using a link structure, video data clips can be held in a directory outside the database, with a reference to their location stored in the database. The kind of database that we needed for the three washback studies was one that could store video information in a systematic way so that a computer program could access the data users required to answer their research questions. The categories available to help a user retrieve relevant video clips would be pre-designed into the system (see below). The data in the database would consist of classroom video-recordings organised according to the categories selected for the analysis of the lessons videoed. The aim was to enable researchers to be able to retrieve classroom action clips useful to their own research through the selection of one or more of these categories.

Classroom video-recording for three impact studies

The classrooms for video-recording in the three impact studies were located abroad as well as in the UK. The more distant recording (in Japan and Cambodia) was done by local participants, who then sent hard copies of their recordings back to us. The videoing in Italy and the UK was planned and carried out by the co-ordinating consultant and Cambridge ESOL personnel, with the advice and support of the Cambridge ESOL Video Unit. To reduce institutional disruption and project costs, just two members of the impact study project team made each of the classroom video-recordings. An easily transportable and relatively simple camera, tripod and microphone were used. Project pilot visits had indicated that sound quality was at least as important as video quality in our investigation of language teaching and learning, especially with group and pair work prominent among classroom activities. A small Panasonic NV-DS27B camcorder was thus selected for all three projects but using a separate directional microphone (AKG semicardoid) on boom rather than the small built-in camcorder microphone. The boom helped the operator to focus on the target groups or individuals and reduce peripheral sound. Camcorder and external microphone were cable-connected; the two-person crew needed to be in constant communication, more often via gestures than microphones and headphones, probably because all concerned were well informed of the purposes of the projects and the videoing. The classroom action was recorded on miniDV cassettes (DVC), each holding one hour of video (11GB of data). For travelling convenience and to reduce intrusion, no additional lighting was used in the classrooms. Our impression was that the participating students and teachers were not significantly affected by the video-recording. An advantage of recorded observation, of course, is that camera shyness, showing off and other possible on-camera effects can be checked for in the repeated replay analyses.

The products of the recordings were in general adequate for analysis, with very little vision or sound data missed. A possible drawback of operating with a minimal video crew, however, is that it makes 'live' note-taking difficult on what is happening in the classroom. This makes it especially important, before and after each video-recorded lesson, to note adequate contextual information: relevant background on the class, teacher and programme; student place chart; learning materials used; lesson plans, if available, and so on (see Plowman 1999, cited above). It also helps if *related* video interviews or focus group discussions with teachers, students, managers and so on, are arranged and recorded during the same period as the classroom videoing.

The Impact Video Database

With 223 classroom video clips so far logged, and an average file-size per clip of around 30 megabytes, our video database occupies nearly seven gigabytes of disk space. The key need was to facilitate researchers' selection from this wealth of classroom observation video data of the clip or clips relevant to their particular research questions. In order to make the clips into which the video data had been divided by the classroom observation analysis identifiable and amenable to selection by the researcher, the categories defining the key characteristics of the clip (i.e. the metadata) needed to be accessible and interactive. Hawkey (2006) gives details of the observation system designed for the three impact studies, its implementation and validation. The categories within the database used to describe each clip (the metadata) were: *date*, *city*, *country*, *timecode*, and then learning/teaching action categories such as *episode*, *activity*, *participation*, *materials* and *comment*. Our focus, as these categories indicate, was on who was participating, where, at what language proficiency level, what kind of teaching/learning activity was being undertaken, as well as, under 'comment', points of interest noted in the original analyses of the recorded classroom observation data. These categories were now the data identifiers, available to assist users to scan the database and then review and analyse the selected video clips. Using such metadata links, the database remains simple and performs well.

The Video Unit project participants were responsible for deciding on software for building the video database. Filemaker Pro, a cross-platform database application from FileMaker Inc., was considered for its broad data type import capacity (e.g. statistical data files, digital pictures, movies etc.) without data entry, and with good searching, sorting, tracking, layout, labelling, reporting and expansion facilities. Microsoft Access was also considered, as a user-friendly database application, convenient for setting up graphical user-interfaces facilitating user manipulation of the contents of the database, which are stored in the background in repository objects called *tables*. The main reason why we finally chose Microsoft Access for the video database was that it was already widely used within the organisation. Technical support needed for the future expansion of the database would thus be available.

Once the classroom lesson videos had been analysed into clips, they were encoded to Mpeg1 files. These combine video and audio files, compressed to manageable sizes for computer use. Each Mpeg1 file has a unique identifier so that it can be linked to an entry in the database. Work now began on creating the user interface (within Microsoft Access) that would enable end-user interaction with the video clips and their associated metadata. The development team noticed that the Access format seemed to enable users to remain in control of what was being displayed on screen and to view the data required with minimal effort.

The Impact Video Database appeared to meet our criteria for the storing and retrieval of the information gathered from classroom observation. We now broadened the database to include video-recorded interviews and focus groups related to the classroom video clips, including conversations with teachers whose classes had been videoed, parents of students in those classes, heads of departments and schools, and programme officials. The interview and focus group clips were programmed into the database with the clip identifying analytic parameters adapted for interview rather than classroom context. The identifying and analysis metadata categories for this sector of the database are: *focus group/interview*, *activity summary* and *main topics covered* rather than the classroom video categories listed above.

The inclusion of the additional data brought a new challenge as some of the *Progetto Lingue 2000* study interview and focus group participants spoke in Italian. In order to make this material accessible for non-Italian speaking researchers, we decided to create English subtitles for the video clips concerned. The research co-ordinator, who had participated in the original interactions, and an Italian-speaking Cambridge ESOL staff member carried out the analysis, translation and editing of the interview and focus group interactions, and agreed on the subtitling for each video sequence. The software to enter the subtitles on to the video clips was SysMedia's WinCaps program combined with their SubtitleScreener package. This process proved fairly straightforward as did the task of integrating the subtitled clips into the database.

The Impact Video Database is intended to facilitate the sharing of research with *bona fide* applied linguistic, language teaching and language testing researchers, for restricted on-site use at Cambridge ESOL. Given the data protection arrangements agreed when the studies were carried out (see Hawkey 2006:91–96), it is necessary for users to meet privacy, confidentiality and intellectual property requirements. Approved users thus sign a research agreement covering uses they may make of the data. This agreement covers, for example, the researchers' obligation to anonymise any reference they may make to the participants or institution of the video lessons they are discussing.[1]

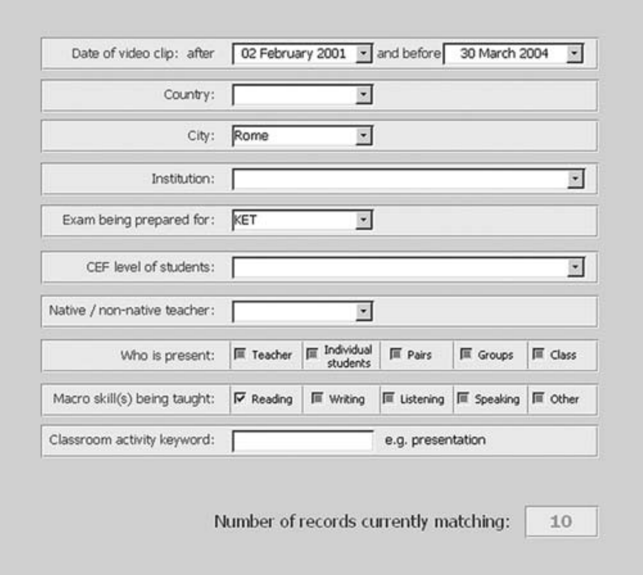
It is, however, envisaged that this video database design will be adapted for use with other video footage already archived at Cambridge ESOL. For example video-recorded samples of spoken English from speakers of different nationalities at various levels of proficiency, could be video-databased. Such clips might well be made available to a global audience through the kind of web-based application mentioned above.

A tour of the Impact Video Database

The Cambridge ESOL Impact Video Database is currently very straightforward in structure, consisting of relatively few screens. Here we take a walk through the database from the perspective of a researcher interested, for example, in the teaching of reading in Rome schools related to Cambridge ESOL's Key English Test (KET), that is, at level A2 of the Common European Framework of Reference for Languages (Council of Europe 2001).

Narrowing the search

With 223 classroom video clips in the database (over 12 hours of footage) the researcher concerned needs to be able to focus on those matching their research needs. On the search screen the range of category options appears to help the researcher to identify the relevant clips in the database. The search screen interrogates the database each time a criterion is updated, thereby allowing the researcher to see if the search is matching any records. For example, entering "Rome" in the city field limits the number of video clips available to 18. The search will then be further restricted by adding 'limiters' such as target English language exam, type of group or skills being taught.



Date of video clip: after 02 February 2001 and before 30 March 2004

Country:

City: Rome

Institution:

Exam being prepared for: KET

CEF level of students:

Native / non-native teacher:

Who is present: □ Teacher □ Individual students □ Pairs □ Groups □ Class

Macro skill(s) being taught: ☑ Reading □ Writing □ Listening □ Speaking □ Other

Classroom activity keyword: e.g. presentation

Number of records currently matching: 10

Figure 1: The search screen showing records matching the criteria: city: Rome, target exam: KET, macro skill: reading

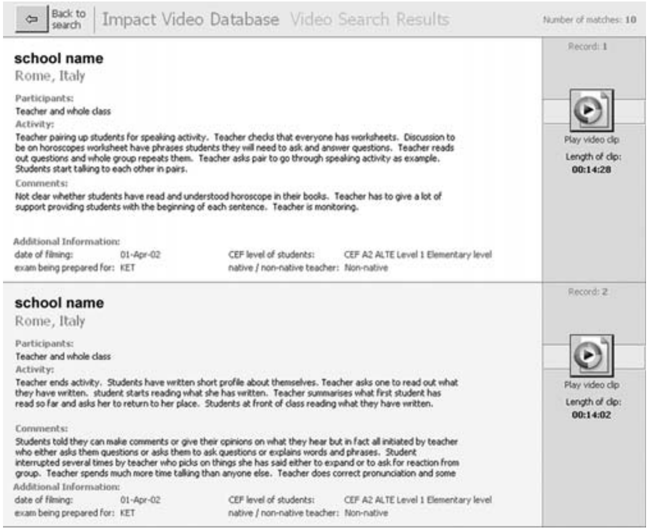
Viewing details

Once researchers have narrowed their search to a more manageable number of clips matching their research needs, they can view the results, as shown in Figure 2. This page displays more detail about each of the clips returned from the search, allowing the user to get a better idea of which clips are most directly applicable to their research questions. Note the buttons on the right hand side of the screen which allow the user to view the chosen video clip.

Watching the clip

The viewing screen (Figure 3) features both the video screen and the details about the targeted clip for reference. The standard Windows Media Player toolbar is also provided beneath the video

[1] Visit our website for details www.CambridgeESOL.org/research



school name
Rome, Italy

Participants:
Teacher and whole class

Activity:
Teacher pairing up students for speaking activity. Teacher checks that everyone has worksheets. Discussion to be on horoscopes worksheet have phrases students they will need to ask and answer questions. Teacher reads out questions and whole group repeats them. Teacher asks pair to go through speaking activity as example. Students start talking to each other in pairs.

Comments:
Not clear whether students have read and understood horoscope in their books. Teacher has to give a lot of support providing students with the beginning of each sentence. Teacher is monitoring.

Length of clip: 00:14:28

Additional Information:
date of filming: 01-Apr-02
exam being prepared for: KET
CEF level of students: CEF A2 ALTE Level 1 Elementary level
native / non-native teacher: Non-native

school name
Rome, Italy

Participants:
Teacher and whole class

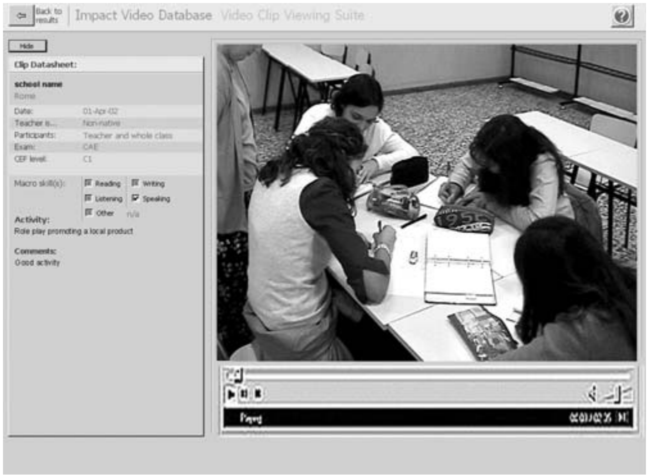
Activity:
Teacher ends activity. Students have written short profile about themselves. Teacher asks one to read out what they have written. student starts reading what she has written. Teacher summarises what first student has read so far and asks her to return to her place. Students at front of class reading what they have written.

Comments:
Students told they can make comments or give their opinions on what they hear but in fact all initiated by teacher who either asks them questions or asks them to ask questions or explains words and phrases. Student interrupted several times by teacher who picks on things she has said either to expand or to ask for reaction from group. Teacher spends much more time talking than anyone else. Teacher does correct pronunciation and some

Length of clip: 00:14:02

Additional Information:
date of filming: 01-Apr-02
exam being prepared for: KET
CEF level of students: CEF A2 ALTE Level 1 Elementary level
native / non-native teacher: Non-native

Figure 2: Details of clips selected from the search



Clip Datasheet:

school name
Rome

Date: 01-Apr-02
Teacher is...: Non-native
Participants: Teacher and whole class
Exam: CAE
CEF level: C1

Macro skill(s): ☐ Reading ☐ Writing ☐ Listening ☑ Speaking ☐ Other n/a

Activity:
Role play promoting a local product

Comments:
Good activity

Figure 3: The viewing screen

screen. This toolbar is a practical set of controls for researchers, enabling them to move to any part of the video clip as required. Of course, users can move back through the screens and continue searching through the clips at their leisure.

Conclusion

The design of an instrument for video data storing and analysis for three related Cambridge ESOL projects has been informed by the extensive educational video use and database literature across the curriculum. The comparative rarity is noted, however, of the kind of critical account attempted here of why and how a practical research video database was developed. The Cambridge ESOL Impact Video Database is already serving its purpose in the analysis of data for the three studies, and has proved robust enough to respond to the data retrieval demands of researchers with specific questions to ask in the context of a broader test validation agenda. The database has also proved amenable to the kinds of modifications and extensions of use mentioned in this article. There is no doubt that, in the washback study of tests on classroom activity, as part of research into the impact and consequential validity of language tests, the use of video databases is demanded. User-friendly models such as the one described here can prove practical and beneficial research tools.

A longer article on the Cambridge ESOL Impact Video Database has been submitted to the refereed journal *Learning, Media and Technology* (previously *Journal of Educational Media*).

References and further reading

Abell, S and Cennamo, K (2003) Videocases in elementary science teacher preparation, in Brophy, J (Ed), 103–129.

Alderson, C and Hamp-Lyons, L (1996) TOEFL preparation courses: A study of washback, *Language Testing* 13, 280–297.

Alderson, C and Wall, D (1993) Does washback exist? *Applied Linguistics* 14, 115–129.

Andrews, S (2004) Washback and curriculum innovation, in Cheng, L and Watanabe, Y (Eds) *Washback in language testing: research contexts and methods*, London: Lawrence Erlbaum Associates.

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

Bailey, K (1996) Working for washback: A review of the washback concept in language-testing, *Language Testing* 13, 257–279.

Benjamin, L N, Brewer, C and Hebl, M (2000) *Handbook for Teaching Introductory Psychology* (Vol 2), Mahwah, NJ: Lawrence W Erlbaum Associates.

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

Brophy, J (Ed) (2003) Using video in teacher education, *Advances in Research on Teaching* (Vol 10), New York: Elsevier Ltd.

Canning-Wilson, C (2000) Role of Video in the F/SL Classroom, in Riley, S, Troudi, S and Coombe, C (Eds) *Teaching, Learning and Technology*, TESOL Arabia 1999 Conference Proceedings, 69–76.

Constantinou, C and Papadouris, N (2004) Potential contribution of digital video to the analysis of the learning process in physics: a case study in the context of electric circuits, *Educational Research and Evaluation* 10 (1), 21–39.

Cornfield, D, Karen, E, Campbell, K, Holly, J and McCammon, H (Eds) (2001) *Working in restructured workplaces: challenges and new directions for the sociology of work*, London: Sage Publications.

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

Dudeney, G (2000) *The Internet and the language classroom*, Cambridge: Cambridge University Press.

Elmasri, N (2004) *Fundamentals of Database Systems*, Harlow: Pearson Education.

FilemakerPro software: www.filemaker.com/products

Fishman, B (2003) Linking on-line video and curriculum to leverage community knowledge, in Brophy, J (Ed), 201–234.

Furth, B and Marques, D (Eds) (2003) *Handbook of video databases: design and applications*, Florida: CRC Press.

Goldman, C, Corley, R and Piaskoski, M (2004) Proceed with caution: the application of antitrust to innovation-intensive markets, *Journal of Information, Law and Technology*, 2004 (1), 1–52.

Green, A (2003) Test impact and English for academic purposes: a comparative study in backwash between IELTS preparation and university pre-sessional courses, unpublished PhD thesis, Centre for Research in Testing, Evaluation and Curriculum in ELT, University of Surrey, Roehampton.

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

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

Hamp-Lyons, E (2000) Social, professional and individual responsibility in language testing, *System* 28, 579–591.

Hawkey, R (2006) *Impact theory and practice: Studies of the IELTS test and Progetto Lingue 2000* (Studies in Language Testing, volume 24), Cambridge: Cambridge ESOL/Cambridge University Press.

Hollingsworth, H (2005) *Learning about teaching and teaching about learning*, paper presented at the Conference of the Australian Council for Educational Research, Melbourne, August 2005.

Institute for Learning Science (2006) *Engines for Education*, Evanston, Illinois: Northwestern University. http://www.engines4ed.org

Jones, N (2001) The role of technology in language testing, *Research Notes* 4, 2–5.

Lowery, L (2005) *The use of multimedia in the teaching of engineering*, presentation to the Department of Civil Engineering, Texas A&M University, February 23 2005.

Maycock, L and Green, A (2005) The effects on performance of computer familiarity and attitudes towards CB IELTS, *Research Notes* 20, 3–8.

Messick, S (1989) Validity, in Linn, R L (Ed) *Educational Measurement* (3rd Ed), New York: ACE/Macmillan, 13–104.

Microsoft Access software: http://office.microsoft.com

Migliacci, N (2002) New ways of using video technology in English language teaching, *ESL Magazine*, May/June.

Plowman, L (1999) Using video for interaction in the classroom, *Spotlight 72*, Scottish Council for Research in Education, www.scre.ac.uk./spotlight/spotlight72.html.

Rhodes, N and Pufahl, I (2003) *Teaching foreign languages to children through video*, www.cal.org/resources/digest/0310pufahl.html

Russow, L (2003) *Digital technology in teaching international business*, Binghampton, New York: Haworth Press.

Saidatul (2006) *Testing spoken English using computer technology*, unpublished PhD thesis, Roehampton University.

Seago, N (2003) Using video as an object of inquiry for mathematics teaching and learning, in Brophy, J (Ed), 259–286.

Shaw, S (2005) Evaluating the impact of word processed text on writing quality and rater behaviour, *Research Notes* 22, 13–19.

Spratt, M (2005) Washback and the classroom: the implications for teaching and learning of studies of washback from exams, *Language Teaching Research* 9 (1), 5–29.

Stempleski, S and Arcario, P (1992) *Video in second language teaching: using, selecting and producing video for the classroom*, Alexandria: TESOL Publications.

Turner, C (2001) The need for impact studies of L2 performance testing and rating: identifying areas of potential consequences at all levels of the testing cycle, in Elder, C, Brown, A, Iwashita, N, Grove, E, Hill, K and Lumley, T (Eds) *Experimenting with uncertainty: essays in honour of Alan Davies* (Studies in Language Testing, volume 11), Cambridge: UCLES/Cambridge University Press, 138–149.

Watanabe, Y (1996) Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research, *Language Testing* 13, 318–333.

Defining the constructs underpinning Main Suite Writing Tests: a socio-cognitive perspective

CYRIL WEIR, CAMBRIDGE ESOL CONSULTANT
STUART SHAW, RESEARCH AND VALIDATION GROUP

Introduction

The credibility of Cambridge ESOL examinations is reliant, to a large degree, upon a coherent understanding and articulation of the underlying latent abilities or construct(s) which they seek to represent. The perceived benefits of a clearly articulated theoretical and practical position for Writing, for example, as a skill area underpinning Cambridge ESOL tests are essentially to deepen understanding of the theoretical basis for how Cambridge ESOL tests different levels of language proficiency across its range of test products; to communicate in the public domain the theoretical basis for the tests; and, to provide a more clearly understood rationale for the way in which Cambridge ESOL operationalise these in its tests.

If the Writing construct(s) is not well defined, then it will be difficult to support the claims Cambridge ESOL wishes to make about the usefulness of its Writing tests, including claims that the tests do not suffer from factors such as *construct under-representation or construct irrelevant variance*. It is widely held that *construct under-representation* (i.e. the test is too narrow in focus and fails to include important elements of the construct of interest) and *construct irrelevant variance* (i.e. a type of systematic measurement error where test score variance is due to factors other than the construct of interest, such as background/cultural knowledge, or unreliable scoring) constitute the two most important threats to construct validity (Davies et al 1999:32–33). Therefore, adequate construct definition for purposes of test validation is a vital principle in language testing.

Having a clear and well articulated position on the underlying construct(s) of Main Suite examinations is also necessary to guide the modifications process for the FCE/CAE Writing tests and to provide a sound rationale for the proposed changes on construct and other grounds such as practicality, impact, etc.

Weir and Shaw (2005) reported briefly on the background to Research and Validation's work in articulating the Cambridge ESOL approach to assessment in the skill area of Writing. To achieve this they developed a socio-cognitive framework, building on Weir

(2005), which views language testing and validation within a contemporary evidence-based paradigm. The framework was used to carry out a comprehensive evaluation of Cambridge ESOL's current approach to examining Writing. This approach has shown itself able to accommodate and strengthen Cambridge ESOL's existing Validity, Reliability, Impact and Practicality (VRIP) approach (see Saville 2003). The new framework seeks to establish similar evidence, but in addition it attempts to reconfigure validity to show how its constituent parts interact with each other.

The results from developing and operationalising the framework with regard to second language writing ability are encouraging, and evidence to date suggests that where it has been applied to other examinations it has proved useful in generating validity evidence in those cases too, e.g., in the International Legal English Certificate (ILEC), the Teaching Knowledge Test (TKT), and BEC and BULATS (see O'Sullivan 2006).

In this article we attempt to summarise our findings in applying the socio-cognitive validity framework to the Cambridge Writing examinations. Much of the substantial validity evidence generated by Cambridge ESOL on its Writing examinations has been brought together in the SiLT volume *Examining Second Language Writing: Research and practice* (Shaw and Weir forthcoming). This has helped clarify a number of areas in examining Writing where further research would be beneficial. We also draw on material from an internal report defining the constructs underpinning Reading, Writing, Listening, Speaking and Use of English within the Main Suite tests (Taylor et al 2006).

The tripartite nature of constructs using the socio-cognitive framework

A pure and comprehensive model of language proficiency remains elusive in theoretical terms; nevertheless, as test developers and providers at Cambridge ESOL we need to have recourse to a reasonably well-informed and coherent model of language proficiency if we are to operationalise aspects of this for practical assessment purposes. Such a model needs to deal satisfactorily with the twin dimensions of: (1) aspects of cognition, i.e. the language user's or test taker's cognitive abilities; and (2) features of the language use context. i.e. task and situation, whether in the testing event or beyond the test. These two dimensions constitute two of the core components within our view of construct definition. Within the specific context of language testing/assessment, which is where we seek to operationalise the theoretical construct, there exists a third dimension which cannot be ignored: (3) the process of marking/rating/scoring itself.

In other words, at the heart of any language testing activity we can conceive of a triangular relationship between three critical components:

- the test taker's cognitive abilities
- the task and context, and
- the scoring process.

These three dimensions – which could also be referred to as cognitive validity, context-based validity and scoring validity – make up a reasonably well-developed conceptualisation of construct validity, which has both sound theoretical and direct practical relevance for Cambridge ESOL. The three dimensions reflect a socio-cognitive perspective on the language testing event and are therefore integral to the socio-cognitive framework for test validation which has emerged recently (see Weir 2005).

By maintaining a strong focus on these three components and by undertaking a careful analysis of our tests in relation to these three dimensions, we should be able to provide theoretical, logical and empirical evidence to support validity claims and arguments about the quality and usefulness of our exams, both at the whole exam and at the skill component level.

The remainder of this article will concentrate on the skill area of Writing. Current thinking and research in the Writing construct is first summarised and then a description of how Cambridge ESOL operationalises the construct is given. The final section describes the criterial distinctions which differentiate the Main Suite Writing test levels. Whilst the process of 'differential analysis' has been completed for the Writing construct, similar analyses are currently underway for Reading, Listening, Speaking and Use of English.

What does the literature say about the construct of Writing?

For many years the notion of writing was decontextualised and regarded primarily as product-oriented, where the various elements are coherently and accurately put together according to a rule-governed system; the text product was seen as an autonomous object and writing was considered independent of particular writers or readers (Hyland 2002). Written products were largely viewed as ideal forms capable of being analysed independently of any real-life uses.

More recently, writing has come to be viewed as a strongly contextualised phenomenon which should not be disconnected from key parameters such as the relationship between the writer and the reader, and the purpose of the writing. According to Hayes (1996:5), writing is fundamentally a communicative act: 'we write mainly to communicate with other humans'; Hamp-Lyons and Kroll (1997:8) offer a similar broad, conceptual view of writing: 'an act that takes place within a context, that accomplishes a particular purpose, and that is appropriately shaped for its intended audience'.

According to this view, the linguistic patterns employed in a piece of writing are influenced by contexts beyond the page which bring with them a variety of social constraints and choices. The writer's goals, relationship with readers and the content knowledge they want to impart are accomplished by the text forms appropriate to that social context. This constitutes a socio-cognitive model of writing as Communicative Language Use which takes into account both internal processing (i.e. cognitive or psycholinguistic) and external, contextual factors in writing. Writing is considered a social act taking place in a specifiable context so particular attention needs to be paid to:

- The writer's understanding of the knowledge, interests and expectations of a potential audience and the conventions of the appropriate discourse community as far as this can be specified.
- The purpose of the writing.

- The writer taking the responsibility for making explicit the connections between the propositions and ideas they are conveying and hierarchically organising their writing.
- The importance of the demands the task makes in terms of language knowledge: linguistic, discoursal and sociolinguistic, and content knowledge.

Research indicates that categories of L2 learners can be differentiated from each other by their age, standard of education, L1 literacy and by their ability and opportunity to write in a second language; the types of writing produced by such groups of L2 writers in real-world language use are also varied. Vahapassi's (1982) General Model of Writing Discourse for describing and categorising writing text types in terms of their most important features suggests that the writing requirements of different groups of L2 learners differ in regard to both cognitive demands and communicative function. These differences are especially important when constructing or developing appropriate tests of writing. A definition of writing ability for a specific context therefore needs to take account of the group of L2 writers identified and the kinds of writing they would typically produce.

Figure 1: Cognitive and Context Validity parameters in Writing (based on Shaw and Weir forthcoming)

COGNITIVE VALIDITY	CONTEXT VALIDITY	
COGNITIVE PROCESSES	**Setting:**	**Linguistic Demands:**
	Task	*Task Input & Output*
• Macro-planning • Organisation • Micro-planning • Translation • Monitoring • Revising	• Response format • Purpose • Knowledge of criteria • Weighting • Text length • Time constraints • Writer-reader relationship	• Lexical resources • Structural resources • Discourse mode • Functional resources • Content knowledge
	Setting:	
	Administration	
	• Physical conditions • Uniformity of administration • Security	

How does Cambridge ESOL operationalise the construct of Writing?

In line with current views on the nature of writing, the model adopted for designing and administering writing tasks in the Cambridge ESOL tests looks beyond the surface structure manifested by the text alone; it regards the text as an attempt to engage the reader communicatively.

The key to defining the relevant Writing construct is in identifying what factors make up the real-world language use, and which of those factors are necessary for what is (and is not) to be measured. Alderson (2000:118) notes that constructs are not so much 'psychologically real entities that exist in our heads', but abstractions defined for a particular assessment purpose. In other words, 'there is no one single definition of language ability that will be applicable for all situations' (Weigle 2002:42). This means that a definition of the construct must be developed for each testing situation and it must take into account the test takers, the purpose of the test and the real-life situation the test is trying to 'simulate'. For this reason, the information gathered about the test-taking populations for our Cambridge tests (e.g. via the Candidate Information Sheets) is of great importance and feeds directly into decisions about test design across the different exam suites and proficiency levels.

Cambridge ESOL follows the socio-cognitive approach in its Main Suite examinations (e.g. for the revision of FCE, CAE) where attention is paid to both context-based validity and to cognitive validity in terms of the cognitive processing and resources that are activated within the test taker by the test tasks (see Figure 1).

Context-based validity involves not just linguistic content parameters, but also the social and cultural contexts in which the task is performed. For a Writing task context-based validity thus addresses the particular performance conditions, the setting under which it is to be performed (such as purpose of the task, time available, length, specified addressee, known marking criteria as well as the linguistic demands inherent in the successful performance of the task) together with the actual examination conditions resulting from the administrative setting. Cognitive processing in a Writing test never occurs in a vacuum but is activated in response to the specific contextual parameters set out in the test task rubric. These parameters relate to the linguistic and content demands that must be met for successful task completion as well as to features of the task setting that serve to delineate the performance required.

The 'symbiotic' relationship between context validity, cognitive validity and scoring validity constitutes what we refer to as *construct validity*. For example, decisions taken with regard to parameters in terms of task context will impact on the processing that takes place in task completion. Likewise scoring criteria where made known to candidates in advance will similarly affect executive processing in task planning and monitoring and revision. The scoring criteria in writing are an important part of the construct as defined by context and processing as they describe the level of performance that is required. Particularly at the upper levels of writing ability, it is the quality of the performance that enables distinctions to be made between levels (Hawkey and Barker 2004). The interactions between and especially within these different aspects of validity may well eventually offer further insights into more closely defined levels of task difficulty.

What are the criterial distinctions between the way Writing is assessed at the five Main Suite levels?

In the forthcoming Writing volume, Shaw and Weir review the latest research and practice in writing assessment, and evaluate Cambridge practice in the light of current theory in applied linguistics and language assessment. Within each constituent part of the validation framework criterial parameters for distinguishing between adjacent proficiency levels in the CEFR are also identified.

Chapter 3 of the volume analyses the cognitive processing that appears to be taking place at the various levels in the Cambridge Main Suite examinations while Chapter 4 examines the ways in which Cambridge Main Suite Writing tests operationalise various contextual variables. Chapter 5 addresses scoring validity which embraces all aspects of reliability and accounts for the extent to which test scores are based on appropriate criteria, exhibit consensual agreement in their marking, are as free as possible from measurement error, stable over time, consistent in terms of their content sampling and engender confidence as reliable decision-making indicators. Aspects of cognitive processing and contextual variables of the task are presented in summary form below.

Cognitive processing

In all Writing tasks at all levels careful task specification (e.g., in terms of purpose, readership, length, known assessment criteria) promotes the stages of macro-planning, organisation, micro-planning, translation, monitoring, and revision.

From PET Part 3 writers are provided with some autonomy and responsibility for shaping and planning the structure and outcome of their discourse. Planning, monitoring and revising written work for content and organisation is increasingly necessary in FCE, CAE and CPE particularly at CAE and CPE levels. From FCE upwards there is a need to engage in knowledge transforming rather than knowledge telling though this is not always required at FCE.

Contextual parameters

We describe below five aspects of the task relevant to context validity.

Response format

KET is characterised by controlled tasks at the word level and limited semi-controlled tasks at the text level. PET Part 1 is controlled, Part 2 and the Part 3 tasks are semi-controlled. At FCE, CAE and CPE there is a mixture of semi-controlled tasks where the task is framed by the rubric and/or input texts but candidates are expected to make their own contribution.

Purpose

There is a transition from KET to CPE in terms of purpose, with the possibility of having to deal with conative purpose (intended to persuade or convince) from FCE level upwards. Only at CPE, however, is the discursive task compulsory. Within the higher levels (FCE, CAE, CPE) the same broad range of purposes for writing may occur at each of the three levels.

Text length

There is in general an increase of about 100 words between each of the first three levels if one takes the minimum amount required as the benchmark. The upper word limit at FCE is substantially greater than that which is expected of KET and PET candidates. There is also substantial difference between the minimum required at CAE and at FCE. Longer pieces of writing will in themselves add to the cognitive pressures on the writer.

Time

At FCE the time available is dedicated time for the Writing tasks alone rather than being shared with the Reading tasks as in KET and PET. There is a substantial increase in the amount of time available at CAE and CPE levels. This increase in time allocation matches the increase in length of writing output.

Writer-reader relationship

There is a gradual progression through the levels from personally known (e.g. friend or teacher) to specified audiences with whom candidates are not personally acquainted (e.g. an editor or magazine readers). Addressing a broader range of audiences is required between PET and FCE as candidates only write to people they know personally in KET and PET. By PET, the candidates also need to take greater account of their audience by considering what the potential reader is likely to know about the subject, the amount of explanation required and what can be left implicit. By CAE, candidates are no longer writing to people they know personally. A slightly wider range of unacquainted audience distinguishes CAE and CPE. At these two levels candidates must decide what sorts of evidence the reader is likely to find persuasive. With the exception of KET, the effect of the writing on the reader is taken into account in the marking.

We now detail the linguistic demands of the task rubric (input) and candidate response (output) that are criterial distinctions between Main Suite levels.

Lexical resources

At the KET and PET levels lexical items normally occur in the everyday vocabulary of native speakers using English. At FCE level topics need to be addressed in more detail and with greater lexical precision. For CAE and above the language expected is more sophisticated and the tasks more lexically challenging than at FCE. Topics, tasks and functions which only require simple language are avoided at the higher levels. At FCE and above there is also an expectation that candidates are able to reformulate input language in their own words. Language associated with conative functions is needed for tasks at CAE and CPE levels.

Structural resources

There is a gradual progression in the complexity of the grammatical constructions required by tasks. This is in line with the structural levels appearing in ELT course books aimed at language levels corresponding to the Council of Europe levels A2 (KET) through to C2 (CPE). At KET level candidates are expected to have control over only the simplest exponents for the Waystage functions at this level. The marker is tolerant of basic errors such

as missing third person ‘s’ and misuse of articles. At PET level candidates show a degree of ability to handle some of the exponents listed at Threshold level. Although the marker is primarily interested in the extent to which meaning is conveyed, control with regard to such basic structures as ‘to be’ agreement is expected. However, in PET Part 3 where candidates demonstrate ambition their writing may be judged adequate even if flawed. At FCE level candidates should have a good grasp of Vantage level language. They should have mastered the main structures of the language and should not be prevented from communicating by a lack of structural resources. As long as the marker does not have to make an effort to understand the writer’s meaning, errors with such aspects of language as gerunds/infinitives or some confusion between the past simple and present perfect will not be unduly penalised. At FCE level candidates tend to write either simply and accurately or more ambitiously but less accurately. Both types of candidates may achieve adequate performance if other aspects of their writing are satisfactory. By CAE candidates are expected to use the structures of the language with ease and fluency. There should be some evidence of range; very simple but accurate language is not enough at this level. Candidates must be able to demonstrate some ability to use complex structures even though they are not expected to write error-free prose. CAE candidates must also show that they have a grasp of structures which allow them to express opinions and feelings in an appropriate register. They can, for example, express dissatisfaction in a manner that does not sound aggressive by using appropriately tentative structures. By CPE level, candidates should demonstrate a high degree of range and accuracy with regard to structures. They should have a mastery of the structures needed to present ideas and attitude in a well-organised and sophisticated manner. Some errors will be tolerated so long as they do not confuse the reader in any way; for example, an inappropriate use of a preposition after a verb or an omitted article will not in themselves cause the candidate to lose marks.

Discourse mode

There appears to be a clear distinction between PET and FCE. At FCE the rhetorical task of argument differentiates it from PET and discursive tasks are important throughout FCE, CAE and CPE. CAE is differentiated from FCE by the greater range of genres the candidate might have to address overall and in the compulsory Part 1 task having to deal with varying degrees of persuasion with the intended audience having to be convinced of the writer's point of view. At CPE candidates might have to write an essay (a genre not previously encountered at lower levels).

Functional resources

There is a clear functional progression across the first three levels (KET, PET and FCE) in terms of complexity but also in the degree of precision in the structural exponents employed to fulfil the function(s). Functions associated with conative purposes and argumentative tasks appear at CAE. The functions at CAE and CPE are increasingly diverse and demanding and intended to produce more complex structures or evidence of collocational knowledge.

Content knowledge

At KET level candidates need to have the language to deal with personal and daily life: basic everyday situations and communication needs (van Ek and Trim 1998a). The focus tends to be on topics that are accessible to teenage candidates. At PET level a broader range of general topics relating to the candidate's personal life and experience is covered; narrative topics also feature at PET level (van Ek and Trim 1998b). FCE candidates may be expected to deal with a wide range of knowledge areas including any non-specialist topic that has relevance for candidates worldwide (van Ek and Trim 2001). CAE candidates are expected to be able to deal with topics that are more specialised and less personal than those that tend to feature at lower levels. The step up to CAE also involves coping with lexically challenging topic areas (e.g. the environment, the scientific world, traditions). At CPE level more abstract and academic topics appear and the candidate may be expected to be able to write on any non-specialist topic. CPE candidates are expected to be able to operate confidently in a wide variety of social, work-related and study-related situations. At all levels topics that might offend or otherwise unfairly disadvantage any group of candidates are avoided.

Conclusion

The issues of what a language construct is and whether it is possible to identify and measure developmental stages leading towards its mastery are critical for all aspects of language learning, teaching and assessment. Exam boards and other institutions offering high-stakes tests need to demonstrate evidence of the context, cognitive and scoring validity of the test tasks they create to represent the underlying real-life construct. They also need to be explicit as to how they operationalise criterial distinctions between levels in their tests in terms of the various validity parameters discussed above.

The Writing volume *Examining Second Language Writing: Research and practice* (Weir and Shaw forthcoming) marks the first comprehensive attempt by any examination board to expose the totality of its practice to such scrutiny in the public arena. Much has already been achieved by Cambridge ESOL and other researchers towards a better understanding of the nature of second language writing proficiency and how it can be assessed; nevertheless, the Writing volume also shows that there are many questions still to be answered and a great deal of work remains to be done. Future research needs to investigate whether further work on refining the parameters discussed in the volume, either singly or in configuration, can help better ground the distinctions in proficiency in writing represented by levels in Cambridge ESOL examinations and their external referent the CEFR, as well as in the level-based tests produced by other language examination boards. This will be a long and challenging road but an essential journey for all of us who are members of the worldwide language testing community.

References and further reading

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

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

Davies, A, Brown, A, Elder, C, Hill, K, Lumley, T and McNamara, T (1999) *Dictionary of Language Testing* (Studies in Language Testing volume 7), Cambridge: UCLES/Cambridge University Press.

Hamp-Lyons, L and Kroll, B (1997) *TOEFL 2000 – Writing: Composition, Community, and Assessment*, TOEFL Monograph 5, Princeton: Educational Testing Service.

Hawkey, R and Barker, F (2004) Developing a common scale for the assessment of writing, *Assessing Writing*, 9 (2), 122–159.

Hayes, R J (1996) A new framework for understanding cognition and affect in writing, in Levy, C M and Ransdell, S (Eds) *The science of writing*, Mahwah, NJ: Erlbaum.

Hyland, K (2002) *Teaching and Researching Writing*, London: Longman.

Milanovic, M and Saville, N (1996) *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem* (Studies in Language Testing volume 3), Cambridge: UCLES/Cambridge University Press.

O'Sullivan, B (2006) *Issues in Testing Business English: the revision of the Cambridge business English certificates* (Studies in Language Testing volume 17) Cambridge: UCLES/Cambridge University Press.

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

Shaw, S D and Weir, C J (forthcoming) *Examining Second Language Writing: Research and practice*, Cambridge: Cambridge ESOL/Cambridge University Press.

Taylor, L, Barker, F, Geranpayeh, A, Green, A, Khalifa, H and Shaw, S (2006) *Defining the construct(s) underpinning the Cambridge ESOL UMS tests: a socio-cognitive perspective on overall language proficiency and the four language skills*, Cambridge ESOL internal report.

Vahapassi, A (1982) On the specification of the domain of school writing, in Purves, A C and Takala, S (Eds), *An international perspective on the evaluation of written composition*, Oxford: Pergamon, 265–289.

van Ek, J A and Trim, J L M (1998a) *Waystage 1990: Council of Europe*, Cambridge: Cambridge University Press.

—(1998b) *Threshold 1990*, Cambridge: Cambridge University Press.

—(2001) *Vantage*, Cambridge: Cambridge University Press.

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

Weir, C J (1988) Construct validity, in Hughes, A, Porter, D and Weir, C (Eds), *ELT Validation Project: Proceeding of a Conference Held to Consider the ELTS Validation Project*, British Council and UCLES internal report.

—(2005) *Language Testing and Validation: An Evidence-Based Approach*, Oxford: Palgrave.

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

Weir, C J and Shaw, S D (2005) Establishing the Validity of Cambridge ESOL Writing Tests: towards the implementation of a socio-cognitive model for test validation, *Research Notes* 21, 10–14.

A worldwide survey of examiners' views and experience of the revised IELTS Speaking Test

ANNIE BROWN, MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH, ABU DHABI, UNITED ARAB EMIRATES
LYNDA TAYLOR, RESEARCH AND VALIDATION GROUP

Introduction

The IELTS Speaking test was redesigned in 2001 with a change in structure and assessment procedure, moving from a semi-structured, 'conversational' style of interview to a more structured format, and from a single holistic band scale focusing primarily on communicative and functional ability, to four analytic scales focusing on linguistic quality. These changes responded to two major concerns; that the lack of consistency in interviewer behaviour could unfairly advantage or disadvantage candidates (Taylor 2000), and that there was an inappropriate level of inconsistency in interpreting and applying the scale (Taylor and Jones 2001).

A number of studies of interview discourse informed the decision to move to a more structured format, including those by Lazaraton (1996a, 1996b) and Brown and Hill (1998). These studies found that despite training, examiners had their own unique styles and differed in the degree of support they provided. Brown and Hill's study, which focused specifically on behaviour in the IELTS interview, indicated that these differences in interviewing technique had the potential to impact on ratings achieved by candidates (see also Brown 2003a and 2005). The revised test was designed with a more tightly scripted format (using interlocutor 'frames') in order to ensure that there would be less variation in the interviewing technique of examiners. A study by Brown (2003b) conducted one year into the operational use of the revised interview found that this was, in general, the case.

In terms of *rating* consistency, a study of rating behaviour on the original IELTS interview (Brown 2000) revealed that whilst examiners demonstrated a general overall orientation to features within the band descriptors, they appeared to interpret the criteria differently and included personal criteria not specified in the band scales (in particular *interactional* aspects of performance, and *fluency*). In addition, examiners appeared to weight different aspects of the criteria differently. Together these led to rating variability. Taylor and Jones (2001:9) report that 'it was felt that a clearer specification of performance features at different proficiency levels might enhance standardisation of assessment'.

Subsequent to the introduction of the revised Speaking test, Cambridge ESOL commissioned a large-scale survey in 2005 to explore examiners' views and experiences of the revised interview. A two-page survey was designed with input from Senior Examiners. It was divided into two sections: *Interview format and tasks*, and *Assessing interview performance*, and consisted of questions requiring both Likert-scale and open-ended responses.

The survey was distributed to examiners in the 30 largest test centres worldwide. A total of 269 responses were received, from 22 centres.

On the whole, the analysis of response patterns to the questionnaire indicated that the interview format and assessment criteria in the revised IELTS Speaking test met with a high level of approval. Overall, ratings were consistently more positive than negative. While it was to be expected that most of the written comments elicited in the questionnaire would focus on the more negative aspects of the test, it was certainly the case that a number of examiners explicitly stated that they felt the interview and the scales worked well. Some commented that they felt the current interview was an improvement on the earlier one, being fairer, easier to manage, and with criteria which reflect candidate ability more accurately. The detailed findings of the study are reported in the remainder of this article.

Findings

Interview format and tasks

In this section of the survey, two areas were covered, *Overall interview format* and *Interview sections*.

Overall interview format

Examiners were asked to respond on a scale of 1 (*Strongly disagree*) to 5 (*Strongly agree*) to seven statements about the overall interview format (see Table 1). They were also invited to include additional written comments elaborating on these ratings. The overwhelming majority of responses were positive on all items, with the percentage of examiners selecting *Agree* or *Strongly Agree* ranging from a high of 92% to a low of 78%.

Table 1: Overall interview format

Statement	N	% selecting Strongly agree and Agree	Mean	S.D.
1 The interview format is appropriate for GT candidates	262	81	4.00	0.80
2 The interview format is appropriate for Academic candidates	264	87	4.17	0.76
3 The interview format is effective in generating assessable discourse	265	85	4.06	0.72
4 The interview as a whole elicits a satisfactory range of language (structures, lexis, functions)	265	90	4.16	0.69
5 There is an appropriate progression in terms of linguistic complexity across the interview	265	87	4.17	0.77
6 The interview format is easy to manage	264	92	4.33	0.70
7 The scripted frames are easy to use	263	78	4.02	0.86

The majority of examiners felt that the interview was appropriate for both types of candidate, although it had slightly less support for General Training candidates. Some examiners expressed concern that General Training candidates might be disadvantaged by the format because they often lacked the maturity to elaborate or the world knowledge or experience to talk about some topics. Other than having a different format for the two types of candidate, one suggestion for dealing with this problem was to allow examiners more freedom to provide cues to elicit elaboration. Interestingly, recent post-test analysis done by Cambridge ESOL to monitor IELTS examiners' take-up of the Speaking tasks available to them showed that the full range of tasks is used with General Training candidates; there was no evidence of examiners avoiding use of specific tasks with this candidate group. In addition, monitoring of candidate performance in the Speaking Test also shows General Training candidates to be coping well with the topic as well as task demands.

The majority of examiners reported that the interview format was effective in generating assessable discourse and elicited a satisfactory range of language responses, although some commented that they felt some topics were culturally inappropriate, either for the particular part of the world, or for rural candidates in some parts of the world.

The lowest level of agreement in this section (78%) was with the statement *The scripted frames are easy to use*. Examiners commented that the wording of some frames or questions was sometimes awkward, over-lengthy or unclear, or that they included colloquial or too culture-specific – British – language. Subsequent comments indicated that this was particularly the case in Part 1 of the interview. Low frequency vocabulary was felt to be a problem for lower-level candidates, for whom many questions were said to be incomprehensible, particularly as examiners were not allowed to explain unknown words or rephrase questions.

Interview sections

Examiners were asked to respond on a scale of 1 (*Strongly disagree*) to 5 (*Strongly agree*) to four statements with reference to each of the three sections of the interview, and to a further statement for each of Parts 2 and 3 (see Table 2). They were also invited to include additional written comments elaborating on these ratings.

Table 2: Interview sections

Statement	% selecting Strongly agree and Agree		
	Part 1	Part 2	Part 3
Part … elicits a sufficient sample of language.	69	86	92
It is easy to manage the timing of Part …	71	89	94
The Part … topics are suitable for the candidature I normally test.	55	70	74
The Part … topics are equivalent in terms of difficulty.	34	39	48
The Part 2 questions provide a suitable bridge to Part 3	-	45	-
I find it easy to tailor the Part 3 prompts to the level of the candidate.	-	-	77

For all three sections, responses to the first two statements were more positive than to the last two. Responses were considerably less positive overall for Part 1 than for the other sections across all statements.

In relation to the first question, some examiners again commented that since the general nature of Part 1 topics was well-known in the public domain, candidates often produced rehearsed responses which were not possible to assess. They also commented frequently on the apparent lack of equivalence or appropriateness of topics. Some Part 1 and Part 2 topics were felt to be inherently more complex than others, some were considered inappropriate for younger candidates, and some were seen as culturally inappropriate. Some Part 2 topics were considered too dull, or too simplistic, for higher level candidates. Examiners commented on the inappropriateness of some Part 3 questions for certain cultures or for younger candidates who had little life experience (e.g. of travel and other cultures). Some topics were described as too abstract or too general, and questions as vague, obscure, or abstract. The suggestion was made to include different Part 2 and 3 frames for adult or professionals and for high school students, or for Academic and General Training candidates. These comments highlight the challenges faced by test producers in producing a Speaking test which is capable of measuring successfully across a broad proficiency continuum and in which the tasks and topics selected need to be accessible to a highly heterogeneous, international candidature. It may be worth noting here that the IELTS test production process requires all Speaking tasks to be trialled on a representative sample of the target candidature before being used in the live test context; this is done to ensure that topics and tasks are maximally accessible to test takers and are capable of generating an appropriate language sample for assessment purposes. Routine post-test analyses of candidate performance in the Speaking test suggest that on the whole candidates cope well with task and topic demands, and that different tasks perform in comparable ways. IELTS examiners sub-select from a range of Part 1 frames, the first of which is designed to be very familiar so that candidates have chance to settle down in the early part of the test; examiners then go on to select from a range of task combinations for Parts 2 and 3 so they do have some flexibility to tailor a topic to an individual candidate. However, a recent internal survey of Part 2/3 task usage showed no clear pattern of examiners targeting specific tasks at particular candidate groups or avoiding other tasks with certain groups. A recent experimental study, which used retired IELTS Part 2 tasks as part of a larger IELTS joint-funded study to investigate task factors, provided valuable evidence of their comparability (see Weir, O'Sullivan and Horai 2006).

A number of examiners commented that the Part 2 rounding-off questions were often inappropriate, either because candidates had already covered the content in their long turn or because they had become irrelevant as the candidate had moved off topic. Some examiners reported that they chose not to ask these questions; others asked them despite knowing that they did not fit. Several requested that the questions be made optional, or that more flexibility be allowed to change them in order to make them more appropriate to the preceding talk. It was also noted that such flexibility would allow examiners to probe responses which were felt to be rehearsed. Interestingly, these findings corroborate results from another recent joint-funded study by Seedhouse and Egbert (2006) and will help to inform possible future changes to the examiner frame.

While three quarters of the examiners felt that it was easy to tailor the Part 3 prompts to the level of the candidate, the reasons given by others for finding this difficult varied. There was a perception that some questions were just too complex, both syntactically and semantically, to simplify and yet retain the general question objective. There was also a perception that the vocabulary was often too difficult for lower level and General Training candidates (e.g. *compensate*, *pace*, *impact*). Two examiners commented that they had not been trained to tailor the prompts; it was not clear whether they were commenting on inadequate focus on this in the training or whether they had specifically been told not to change the wording.

More positively, examiners commented that allowing them to tailor the prompts allowed them to cater to candidates dealing with topics differently. A number of examiners commented positively about the value of Part 3 in eliciting a good language sample, and allowing assessment of higher level candidates. Examiners in some regions of the world were concerned that this section was the only one where candidates could not memorise answers, and was thus the only part they could assess for some candidates.

Assessing interview performance

Examiners were asked to rate each of the four scales with regard to ease of interpretation, ability to discriminate between levels of proficiency, and confidence in the accuracy of their ratings. For each statement there was an over 80% agreement rating on all scales except *Pronunciation*, which scored considerably lower (see Table 3).

Table 3: The four scales

Statement	% selecting Strongly agree and Agree			
	F&C	LR	GRA	PRO
The descriptors are easy to interpret	85	85	89	66
The descriptors discriminate clearly between levels of proficiency.	80	80	81	59
I feel confident that my ratings are accurate when applying the scales.	91	86	89	71

For all scales, some examiners commented that that the terminology used in the descriptors was subjective, vague or otherwise problematic to interpret. They referred to the terms such as *sufficient*, *limited*, *basic*, *effective* and *occasional*, *wide range*, etc. There were also comments about the difficulty of distinguishing particular adjacent bands, although which were the most difficult to distinguish varied from examiner to examiner.

In relation to the *Fluency and Coherence* scale, a number of examiners remarked that the criteria did not allow them to take into account that a candidate might be fluent but speaking off-topic. Others commented that it can be difficult to determine whether hesitation is caused by a search for lexis or grammar, or content. Assessments of the speed of speech were felt to be problematic for candidates who spoke fluently but too fast, or started slowly but then kept going, and for those who spoke slowly because they deliberated over what they were saying. Speech that was coherent but short was also reported to be problematic to assess. There was a perception that the Fluency and Coherence scale was the most subjective. The bands most commonly described as the most difficult to distinguish were 5 and 6, and 8 and 9.

In relation to the *Lexical Resource* scale, the role of word forms or parts of speech was reported to be unclear, as was how to deal with candidates who use high level vocabulary inappropriately, or those who produced sophisticated vocabulary in rehearsed responses. Levels 4 and 5, and 5 and 6, were reported as the most difficult to distinguish.

Difficulties reported when assessing *Grammatical Range and Accuracy* included that one can become attuned to errors, or (conversely) one can be too critical of errors; it is difficult to keep track of errors and complexity, especially when participating in the interaction; it is difficult as a non-native speaker of English to know what "native speaker errors" are. Bands 5 and 6 were reported as the most difficult bands to distinguish.

Pronunciation was rated the lowest on all three statements. A considerable number of examiners commented on the lack of distinction offered by four *Pronunciation* bands. Some remarked that having to award a 4 or a 6 could skew the overall result, and some reported that because of this they took the impact of the *Pronunciation* rating on the overall score into account before deciding what rating to award. In response to a subsequent question which asked *Which of the scales are you least confident about?*, by far the largest proportion of examiners selected *Pronunciation* (see Table 4). There was a strong feeling that additional (i.e. the odd-numbered) bands were necessary.

Table 4: Confidence in the scales

	F&C	LR	GRA	PRO
Which of the scales are you least confident about?	20	13	15	52

Several examiners reported that they found it difficult to assess *Pronunciation* because they felt they were attuned to certain types of pronunciation. Conversely, they felt they could be too critical of the pronunciation patterns in English of speakers of their own language. A number of comments related to the question of what "native speaker" speech was and what sorts of regional variations were acceptable (for example, African English, Asian English). Clearly enunciated but monotone speech was also felt to be problematic to assess. Issues of familiarity with certain types of pronunciation and of the impact of regional variation are important dimensions of the assessment process that are relatively under-researched; nevertheless, given their direct implications for speaking assessment, such issues are being increasingly discussed in the literature (see Taylor 2006) and are attracting growing attention from researchers.

In response to a request for comments on the assessment process in general, the non-compensatory nature of the scale, that is, the requirement that the performance must fulfil all aspects of the descriptor to be assessed at a particular level, was disliked by some examiners. This was perceived to be a concern particularly in relation to *Pronunciation*, where there was a large difference between bands. Other examiners reported that they felt the need for a way of dealing with a lack of comprehension of prompt (e.g. by assessing comprehension), or when candidates go off-topic (see comments on the *Fluency and Coherence* scale, earlier). The suggestion was made that a task-response scale could be included or that, as in the Writing test, there should be a specific focus on how the candidate addresses the question. A desire to award half-bands or to have overall Speaking test performance reported in half-band increments was also noted. The possibility of introducing half-bands for reporting IELTS Speaking and Writing results has been under consideration for some time; following successful outcomes from a number of studies to explore the impact of such a change, a decision has been made to report Speaking and Writing scores using half-bands from July 2007.

Overall comments on the IELTS Speaking test

Two final questions on the survey asked examiners what aspect of the Speaking test they felt the most and the least confident about. Certain trends were noted, for example that many examiners were most confident about conducting the interview and following the script, whereas many were least confident about making accurate assessments. One aspect of the test examiners reported being most confident about was being able in Part 3 to formulate questions to guide discussion and elicit long responses and discover the candidates' abilities or ceiling. Part 2 was also nominated, as the monologue response allowed examiners to simply listen to the candidate. It was described by one examiner as 'well designed and provid[ing] good scope for accurate assessment as individual strengths and weaknesses are easily revealed.'

More generally, a number of examiners commented that the interview as a whole worked well. The following comments were typical:

- 'The process is clear and fair to all candidates if correctly administered.'
- 'The test flows together well and I have some choice in topics. The testing rubrics are well designed.'
- 'Although I was doubtful at the beginning I now feel that the formal regimentation of the test is very fair and leads to the fairest assessment possible.'
- 'The test is thorough.'

When asked what aspect of the interview they felt least confident about, the following were commented on the most: dealing with candidates who do not understand the prompt, answer off-topic, say very little, or produce rehearsed responses in Parts 1 and 2; whether or how much to deviate from set questions; dealing with really weak candidates in Part 3, and dealing with very proficient candidates in Part 1; being able to stretch candidates sufficiently in Part 3; the imposed 'neutrality' of the examiner, which was felt to be anxiety-provoking for some candidates; and the cultural relevance of some topics. Cambridge ESOL is responsible for producing training and standardisation materials for IELTS examiners, as well as other materials to develop their skills and expertise over time and to support them in their examining work on an ongoing basis through the IELTS Professional Support Network. A recently developed Focus on Procedure video, for example, can be used by IELTS examiner trainers to focus on some of the problems highlighted above and to give them appropriate strategies for addressing these with confidence.

Conclusion

On the whole, the analysis of response patterns to the statements about the interview format and assessment criteria indicated that the revised IELTS Speaking test met with a high level of approval.

The major concerns of examiners participating in this study, that is those commented on the most frequently, were:

- the lack of appropriateness of some topics for both General Training and Academic (younger/more mature) candidates and for different cultures
- a lack of equivalence across topics
- the lack of flexibility in wording prompts and modifying content of prompts
- rehearsed speech and familiarity with test topics, especially in certain regions
- a need to take rehearsed speech and off-topic responses into account when making assessments
- a lack of specificity in the wording of the scales
- a need for greater discrimination in assessing pronunciation.

The large-scale survey of IELTS examiners reported here complements a number of other studies conducted under the Joint-funded IELTS Research Program (see Nicholson and Walsh 2006); together these form part of the ongoing research and validation agenda for the revised IELTS Speaking test introduced in 2001. These studies provide important empirical evidence that the revised test is functioning well overall and as the test developers intended; the findings are also valuable in generating information which can feed directly into examiner training programmes and in highlighting issues which may need attention in any future revision.

References and further reading

Brown, A (2000) An investigation of the rating process in the IELTS Speaking Module, in Tulloh, R (Ed) *Research Reports 1999*, volume 3, Sydney: ELICOS, 49–85.

—(2003a) Interviewer variation and the co-construction of speaking proficiency, *Language Testing*, 20, 1–25.

—(2003b) *A cross-sectional and longitudinal study of examiner behaviour in the revised IELTS speaking test*, unpublished report submitted to IELTS Australia.

—(2005) *Interviewer variability in oral proficiency interviews*, Frankfurt: Peter Lang.

—(2006) An examination of the rating process in the revised IELTS speaking test, in Nicholson, P and Walsh, S (Eds) *IELTS Research Reports*, volume 6, Canberra, Australia: The British Council and IELTS Australia, 41–70.

Brown, A and Hill, K (1998) Interviewer style and candidate performance in the IELTS oral interview, in Woods, S (Ed) *Research Reports 1997,* volume 1, Sydney: ELICOS, 1–19.

Lazaraton, A (1996a) A qualitative approach to monitoring examiner conduct in the Cambridge assessment of spoken English (CASE), in Milanovic, M and Saville, N (Eds) *Performance Testing, Cognition and Assessment: Selected papers form the 15th Language Testing Research Colloquium* (Studies in Language Testing volume 3), Cambridge: UCLES/Cambridge University Press, 18–33.

—(1996b) Interlocutor support in oral proficiency interviews: The case of CASE, *Language Testing* 13, 151–172.

Nicholson, P and Walsh, S (2006) *IELTS Research Reports*, volume 6, Canberra, Australia: The British Council and IELTS Australia.

Seedhouse, P and Egbert, M (2006) The interactional organisation of the IELTS speaking test, in Nicholson, P and Walsh, S (Eds) *IELTS Research Reports*, volume 6, Canberra, Australia: The British Council and IELTS Australia, 161–205.

Taylor, L (2000) Issues in speaking assessment research, *Research Notes* 1, 8–9.

—(2006) The changing landscape of English: implications for language assessment, *ELT Journal*, 60, 51–60.Taylor, L and Jones, N (2001) Revising the IELTS speaking test, *Research Notes* 4, 9–11.

Taylor, L and Jones, N (2001) Revising the IELTS speaking test, *Research Notes* 4, 9–11.

Weir, C, O'Sullivan, B and Horai, T (2006) Exploring difficulty in speaking tasks: an intra-task perspective, in Nicholson, P and Walsh, S (Eds) *IELTS Research Reports*, volume 6, Canberra, Australia: The British Council and IELTS Australia, 119–160.

The effect of editing on language used in FCE Reading texts: a case study

GLYN HUGHES, ASSESSMENT AND OPERATIONS GROUP

Introduction

This article is based on research recently completed as part of an MSc in TESOL at Aston University, UK. The study compared a corpus based on the texts from 13 FCE Reading texts (analysed using Wordsmith Tools, Scott 1999) with the British National Corpus (Leech, Rayson and Wilson 2001; see BNC online website) and subsequently with concordance lines taken from the COBUILD Concordance sampler (see COBUILD website).

Corpus linguistics has increasingly found evidence to back up Sinclair's *idiom principle* for language processing, which argues that language consists of a number of preferred phraseologies that language users utilise in order to encode and decode text (cited in Hunston 2002:143–5). For the FCE Reading test to be a fair assessment of a candidate's ability to decode English, the words in the texts should appear as part of these preferred phraseologies. The purpose of this study therefore was to examine the impact that the editing of texts has on the language used in FCE Reading texts. Having investigated differences in the frequency of various words, the study focused on the extent to which the usage of one word – *people* – differed in FCE and general usage.

Rationale for a case study approach

Comparing the frequency lists from BNC and FCE Reading texts (henceforward *FCE corpus*) corpora, there were very few nouns (other than pronouns) in either list. The first noun to appear in both lists was *people*, although it appears considerably higher in the FCE list (position 42) than in the BNC list (position 86). A comparison of relative frequencies also reveals that *people* is almost three times as frequent in the FCE corpus (3,229 times per million words) as in the BNC corpus (1,146 times). Apart from the possibility of *people* being used in place of less frequent and less widely applicable synonyms such as individuals or inhabitants, it is difficult to see a reason for this.

People would usually be described as a lexical rather than a grammatical word as it does not form part of a closed set of grammatical words. However, it is unlikely to be limited to any particular frame of reference. It could be described as an extreme version of an 'outsider' word, in that it is difficult to imagine insider knowledge of a topic area or genre being necessary to interpret it. For these reasons, it is difficult to attribute the relative frequency of *people* to text selection. This fact in itself made *people* an interesting word to study in detail.

Table 1: Modified nominal groups in two corpora

	BNC Number	FCE Number
unmodified	34	43
pre-modified	40	41
post-modified	13	14
pre- and post-modified	13	9
Total	**100**	**107**

Methdology

This study used a random sample of 100 concordance lines from the BNC written corpus and compared it to the 107 examples of *people* from the FCE corpus. The analysis of the concordance data starts with an analysis of form, specifically looking at the nominal group that contains the word *people* and how it is modified. It then considers the nominal group's position in the clause and in the wider discourse. It also looks at what senses of *people* are used in the two corpora and how these senses relate to the form. For the purpose of clarity, this article uses the terminology of Systemic Functional Grammar to analyse the concordance lines (see Bloor and Bloor 1995).

Findings: the nominal group

In the BNC sample 34 occurrences of *people* are unmodified, 53 are pre-modified and 26 are post-modified in some way, with some instances (13) being both pre- and post-modified. The numbers in the FCE corpus are, given the sample size, very similar. Of 107 occurrences, *people* is unmodified 43 times, pre-modified 50 times and post-modified 23 times. The number and percentage of each type is shown in Table 1.[1]

People pre-modified

Looking in more detail at the manner in which *people* is pre-modified again reveals many similarities between the two sets of data, as shown in Table 2. Rather than using parts of speech, it was considered more appropriate to analyse the pre-modifiers using their *experiential* function (Bloor and Bloor 1995, 137).

There are, however, some differences between the two samples. The FCE corpus contains a much broader range of numeratives than can be found in the BNC. It could be argued that this places an unfair burden on candidates by requiring them to recognise a wider range of numeratives more frequently than they appear in 'the real world', although more evidence would be needed to back up such a claim. Looking at deictics, the large proportion (33%) of occurrences of *other* before *people* in the FCE corpus is a statistical

[1] Note that some concordance lines contain more than 1 example, hence numbers are higher than 100.

Table 2: Functions of pre-modified nominal groups in two corpora

	BNC		FCE	
numerative	16	numbers (6) *many* (5) *more* (2) *all, few, most* (1 each)	19	*many* (5) *a lot of* (4) *most* (3) numbers (2) *lots of, loads of, the majority of, hundreds of, most of the* (1 each)
deictic	19	*the* (9) *his, other, these* (2 each) *Joseph's, some, such, those* (1 each)	21	*the* (12) *other* (7) *some, some of the* (1 each)
epithet	15	nationalities (4) *black, creative, decent, faceless, homeless, old, older, real, small, wealthy, young* (1 each)	12	*young* (3) *local* (2) *right-handed, like-minded, new, retired, small, successful, tanned* (1 each)
classifier	4	lines 30, 54, 61 and 62	1	*record company* (1)

anomaly due to a text about a book called *Other People's Shoes*. The BNC corpus contains no epithets that are used more than once. The FCE corpus contains 3 examples of *young people* and 2 of *local people*.

A further analysis of common collocates of *people* using t-scores on the COBUILD Concordance and Collocations Sampler found that *young* is the ninth most significant collocate for *people* and *local* is the fifty-sixth most significant. This suggests that, although these common collocates may appear more frequently in the FCE texts than they do in a sample of a general corpus, the evidence supports the fact that the phrases are part of preferred phraseologies.

People post-modified

Turning to the manner in which *people* is post-modified, the overall picture once again shows many similarities. Although some different prepositions are used, it is not possible to draw firm conclusions about most of them due to the small sample size and the relative infrequency of prepositional phrases as post-modifiers in either sample.[2]

Once again, however, there are some points of interest worth noting. The BNC sample contains two examples of the phrase *people like*. On both occasions it is used to move from the general to the specific in order to exemplify something, for example: '…before television offered an alternative route to aspiring filmmakers, people like Anderson, Tony Richardson and Karel Reisz worked instead within sponsored…'. This quite complex piece of language is not present in the FCE corpus. Similarly, the FCE corpus

Table 3: Post-modifiers of *people* in two corpora

	BNC		FCE	
relative clause	14	*who…* (11) *whom…, whose…*, no pronoun (1 each)	11	*who…* (11)
…ing verb	4	*working, trying, reaching out, voting*	5	*listening, sitting, climbing chatting, booking lessons*
like…	2		0	
with	2		0	
over	1		0	
of	0		3	age, ability, race and ability (1 each)
in	1	location	3	age, location, profession (1 each)
on	1	location	1	location
other	1	*next door*	1	*older than…*

[2] Examples of *people* followed by a prepositional phrase have only been included where the prepositional phrase is post-modifying the noun, not where it is acting as a clause adjunct.

contains three examples of the phrase *people of*, for example '…providing access to music in a friendly, non-competitive environment, for people of mixed abilities and of all ages and social backgrounds…' whereas this phrase is not found in the BNC sample. In a 40-line concordance sample from the written corpus of the COBUILD Bank of English, the phrase *people of* was used to describe geographical origin 28 times. Of the remaining, two were part of other patterns (e.g. *deprive people of*), one was a false start and only nine were used in a similar fashion to the examples in the FCE corpus (i.e. a non-geographic subset of people). Furthermore, of these nine, three were part of the fixed phrase *people of all ages*. Overall, there is some evidence to suggest some overuse of an unusual pattern but the use of *people of* in this way does occur in written texts in general corpora.

Senses of *people*

We will now look beyond the nominal group and consider the sense of the word *people* in the corpus evidence. In doing so, we will also examine the unmodified examples of *people*. Looking first at the 100 concordance lines sampled from the BNC, it seems that it is possible to divide the manner in which *people* is used in the text into two (or possibly three) different senses. In the 66 of the 100 lines where *people* is either pre-modified, post-modified, or both, the word is being used as a means of clarifying who is being referred to; the writer is using a nominal group including *people* to define who they are writing about, for example: 'business people who are necessary to the profession'. This can even be said of *some people*, where the writer is pointing out that they are not referring to everybody. It is possible to separate this sense further by saying that in some cases (in the BNC sample when talking about nationality), the word *people* is used as part of an assertion of collective identity.[3]

The FCE corpus does not contain any examples of *people* that can be interpreted as an assertion of identity. However, in the other instances where *people* is modified it is performing the same function as in the BNC corpus: it is a useful means of defining who is being referred to, for example 'right-handed people'.

Where *people* is unmodified it is used in a noticeably different manner. Here it is not defining who is being spoken about; instead it is being used as a *denotatum* (Singleton 2000:64) to denote a very general sense of *people*. The people themselves are not important on these occasions; they merely fill a slot in a piece of discourse that is focused on someone else. This general impression is backed up by the fact that *people* is not usually the subject of a clause other than a relative clause. On the few occasions where *people* is the subject of a clause, it is not usually the topical theme of an independent clause. In fact, on the two occasions when it is, the language has unusual features. It is either part of a list of occurrences or used as something of a rhetorical device – a new sentence being used instead of a relative clause.

Again, usage in the FCE corpus is very similar. There are 34 lines where *people* is unmodified and is used denotationally, for example 'if your presence in a film causes people to get into their cars and go to a movie'. There are interesting examples of adjuncts added at the beginning of sentences to prevent *people* from being the topical theme, such as: '*Here* people tended to eat with one eye on their watches …'. However, there are two examples where *people* starts a sentence, for example: 'People ask me if downhill racing is really scary.' On these occasions it is not really possible to explain this in terms of an infrequent rhetorical device so it could be said to be a little unusual.

Conclusion

The analysis of concordance lines for *people* lends support to the idea that differences in frequencies between the three corpora studied are as much to do with text selection as to do with how the language is edited. The analysis furthermore showed little evidence that *people* is being used differently in FCE Reading texts from how it is used in two general corpora of English. Therefore, there is some evidence that candidates should be able to decode the reading texts using the preferred phraseologies that they have encountered in everyday usage. This finding is backed up by the success of candidates in the Reading part of FCE.

It is valuable to find new ways of checking that the content of tests is directly relevant to or closely reflects the world beyond language tests and this sort of corpus analysis is one way of doing this. Through corpus-based studies such as this, Cambridge ESOL staff triangulate corpus and other research methodologies which can then be used to inform test validation procedures.

[3] Many languages – including Slavonic languages – use a different word for this sense of *people*.

References and further reading

Bloor, T and Bloor, M (1995) *The Functional Analysis of English: A Hallidayan Approach*, London: Arnold.

Ghadessy, M, Henry, A and Roseberry, R L (Eds) (2001) *Small Corpus Studies and ELT: Theory and Practice*, Amsterdam/Philadelphia: John Benjamins B.V.

Hunston, S (2002) *Corpora in Applied Linguistics*, Cambridge: Cambridge University Press.

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

Scott, M (2001) Comparing corpora and identifying key words, collocations, and frequency distributions through the *WordSmith Tools* suite of computer programs, in Ghadessy et al.

Sinclair, J (2001) Preface, in Ghadessy et al.

Singleton, D (2000) *Language and the Lexicon: An Introduction*, London: Arnold.

Tognini-Bonelli, E (2001) *Corpus Linguistics at Work*, Amsterdam/Philadelphia: John Benjamins B.V.

UCLES (2005) *FCE Handbook for teachers*, Cambridge: Cambridge ESOL, www.CambridgeESOL.org/exams/fce.htm

Corpus software

BNC online: SARA version 0.941 © Chancellor, Masters and Scholars of Oxford University 1995–97: www.natcorp.ox.ac.uk/tools/sara/index.xml

COBUILD Concordance and Collocations Sampler: www.collins.co.uk/Corpus/CorpusSearch.aspx

WordSmith Tools Version 3.00.00 © Mike Scott 1999: www.lexically.net

Conference reports

The past year has seen participation by Cambridge ESOL staff in a variety of regional, national and international events.

IATEFL 2006, Harrogate, UK

A large number of Cambridge ESOL staff and consultants attended and presented at the 2006 conference of the International Association for Teaching English as a Foreign Language (IATEFL), held this year at the Harrogate International Conference Centre. Representatives from Cambridge ESOL attended pre-conference events organised by the Special Interest Groups (SIGs) for: Young Learners; Testing, Evaluation and Assessment; ES(O)L Teacher Trainers and Educators. Conference papers were given by Sharon Harvey, Lynda Taylor, Martin Robinson, Paul Seddon, Rod Boroughs, Juliet Wilson, Mary Spratt, Heather Daldry, David Clark, and Jacky Newbrook; their presentations covered a wide range of issues related to Cambridge ESOL's approach to assessment, including IELTS, Young Learners, TKT, BULATS, Skills for Life, the Common European Framework, and national education projects. Cambridge ESOL was pleased to be able to sponsor the opening welcome reception for the conference on behalf of the IELTS partners; in addition, over 300 guests attended a special reception held towards the end of the 5-day conference to launch the English Profile project; see www.EnglishProfile.org for further information.

BAAL Corpus Linguistics SIG Seminar, Milton Keynes, UK

A BAAL Corpus Linguistics Special Interest Group (SIG) event was attended by Fiona Barker at the Open University in April 2006. Entitled 'Text analysis using corpora – methodological issues', this one-day event included two plenaries and six papers around the themes of corpora and discourse analysis, especially looking at problems with data identification and issues surrounding spoken data. Around 50 participants took part. Plenary speaker Professor Susan Hunston (University of Birmingham) opened the event with her talk on 'Text and Intertextuality: Debating the issues', followed by Professor Guy Cook (The Open University) whose plenary was entitled '"It just says 'could'. Yes I just spotted that." Corpus facts in discourse analysis'. Speakers in the session on 'Problems with Data Identification' discussed ways of identifying metaphor, phraseology and keywords whilst in the final session – 'Issues around Spoken Data' – speakers concentrated on the technological and analytical challenges facing developers and users of spoken corpora. Visit the BAAL Corpus Linguistics SIG website www.corpus-sig-baal.org.uk for further information.

EALTA Conference, Krakov, Poland

The third annual conference of EALTA (European Association for Language Testing and Assessment) was held in Krakow between 19–21 May on the theme 'Bridging the gap between theory and practice'. The conference included paper, poster and work-in-progress presentations. EALTA also took steps towards creating a code of practice system by first ratifying a code of practice and then debating the issue of what further developments would be required to see the code implemented by members. Much of the debate centred on the difficulty teachers perceived in seeing the code of practice implemented in their workplaces.

Cambridge ESOL and other ALTE members were well represented at the EALTA conference. Nick Saville presented a paper on 'Investigating the impact of language assessment systems within a state educational context', describing how assessment influences a much wider sphere than is often imagined. He looked in particular at the Asset Languages project in the United Kingdom; see www.AssetLanguages.org for further information.

David Thighe presented a paper entitled 'The International Legal English Certificate: Issues with developing a test of ESP' outlining the work being done to ensure that ILEC successfully tests English in a legal context. A principal focus of this work is to verify the correct placement of ILEC on a specificity continuum ranging from *highly specific* (e.g. a test with a great deal of context-related language, such as a test for air traffic controllers) to *general* (e.g. a test for general purposes, like FCE).

Michael Corrigan gave a presentation entitled 'Putting the ALTE Code of Practice into practice: auditing the quality profile'. He described an auditing system which aims both to establish whether testing bodies meet minimum quality standards and to assist them in improvements to their test provision. The presentation focused on the difficulties in constructing such a system and future challenges in fully implementing it. Michael was also able to share the results from the piloting of the system which show that it is already benefiting ALTE members in quality management within their organisations.

David Thighe and Tamsin Walker presented posters at EALTA. David's poster set out Cambridge ESOL's Skills for Life examinations, a recently developed suite of modular tests designed for speakers of English as a second language who are resident in Britain. Tamsin's poster focussed on Asset Languages, showing the progression from the DfES' Languages Ladder statements to the suite of External and Teacher assessments produced by Cambridge ESOL/OCR, and mentioning issues concerned with scale construction, standard setting and test validation.

IVACS Conference 2006, Nottingham, UK

The third Inter-Varietal Applied Corpus Studies Conference (IVACS) was held in Nottingham in June. Fiona Barker attended this event which took the theme of 'Language at the Interface' and over one hundred delegates took part. Four keynote sessions each included two speakers providing alternative, mostly complementary, viewpoints on important issues within corpus linguistics research. Srikant Sarangi and Chris Candlin spoke about 'Aligning research and practice in professional discourse: the case for Case Studies' whilst Martin Wynne and Peter Stockwell presented 'Corpus Stylistics: A Public Enquiry?'. Susan Hunston and Paul Thompson's keynote session was entitled 'Enabling language learning through corpora' and in the final keynote, Anna Mauranen and John Sinclair presented a new analytical procedure, in their talk entitled 'From Text to Tree: LUG, LUM and PUB'.

The conference included papers on a wide range of applied corpus studies, some of which were of particular relevance to language teaching and testing, including those on specific corpora (of business texts, contemporary speech and student writing) and vocabulary studies such as vocabulary difficulty for EFL learners in specific countries and the grading of lexical chunks. Various forms of phraseological analysis were reported in several papers, including the study of multi-word units, stance bundles and evaluative language in academic texts and discourse. Other interesting papers presented lesser-studied features of speech including back-channels and pauses and subjects such as improving the communication of multi-lingual airport groundstaff and a spoken corpus of British working men's speech collected in the 1930s using transcription and note-taking. The third IVACS conference was a successful event which highlighted the growing range of applications of corpus-informed research happening worldwide.

Language Testing Research Colloquium 2006, Melbourne, Australia

Members of Cambridge ESOL staff attended and presented at the 2006 LTRC held this year in Melbourne, Australia in late June. Nick Saville presented a paper entitled 'A model for investigating the impact of language assessment within a national educational context' and Mike Milanovic participated in an invited symposium which discussed 'The social responsibilities of language testers'. Hanan Khalifa and Nick Saville also presented a poster entitled 'Helping ESOL teachers transcend borders: the case of the Teaching Knowledge Test'. At this year's conference banquet, Cambridge ESOL was involved in the presentation of two formal awards. The first of these was the IELTS Masters Award presented to Fumiyo Nakatsuhara on behalf of the IELTS Partners; an article based on Fumiyo's award-winning masters dissertation appeared in *Research Notes* 25.

IELTS Partner representative Anne-Marie Cooper (IDP Education Australia) with IELTS Masters Award recipient Fumiyo Nakatsuhara.

James Purpura (ILTA Vice-President) and Mike Milanovic (CEO, Cambridge ESOL) present Charles Stansfield with John Clark's UCLES/ILTA Lifetime Achievement Award.

The second award was the UCLES/ILTA Lifetime Achievement Award, presented this year to Dr John L D Clark; as John could not be present on this occasion due to a prior commitment, the award was received on his behalf by fellow language tester Charles Stansfield. John's acceptance speech is reproduced below.

John Clark's acceptance speech

I'm certainly delighted, honored, and humbled to be the recipient of the lifetime achievement award, and would like to sincerely thank ILTA, UCLES, and all good friends and colleagues for this totally unanticipated but very much appreciated recognition. I'm told that two minutes, plus or minus 15 seconds, is usually about the upper limit of audience attention for "thank-you-for-this-award" remarks. Nonetheless, in faithful adherence to longstanding LTRC discourse principles and operational practice, I'd like to take at least a bit of this short time to pose two very important questions:

First, how many fledgling statisticians does it take to change a light bulb? Thirty is the absolute minimum……but more experienced practitioners can often get by with only eight or so.

Second, how do linguists know when it's time to retire? When the word "hair" changes from a mass noun to a count noun.

But in all seriousness, folks, what summary thoughts on this occasion might I offer to best convey some of the major "lessons learned" in the course of thirty-plus years' involvement in foreign/second language test development, research, and practical application of testing results? A good number of suggestions come to mind, but time constraints will necessarily limit these to three.

The first observation is the imperative need for test developers to diligently follow what I call the "80/20" rule in carrying out their work. This is the notion that 80 percent of the total test planning and test design effort should be devoted to discussing, defining, and thoroughly explicating the *intended measurement goals* of the test – in other words, to clearly specifying what it should be possible to assert with respect to particular language comprehension or production abilities on the part of the examinee as a result of the testing. This question should be asked and definitively answered *before* the developers even begin to think about testing formats, item types or other technical aspects. It's been my experience that if the "what" of the testing endeavor is thoroughly and unambiguously established ahead of time and in sufficient detail, the appropriate procedural "how" will fall into place quickly, easily, and virtually of its own accord. Conversely, development efforts that launch precipitously into questions of format, item types, and other test-process details without having done the crucial "what are we trying to measure?" spade-work may be expected to become engrossed in unproductive, round-robin debate and discussion of technical issues, largely uninformed by any overarching and guiding concept of measurement purpose.

I'll preface my second suggestion by quickly confirming that some of my best friends are statisticians, and by heartily acknowledging that appropriate statistical analyses and reporting procedures can add considerable informational and practical value to an *inherently good test* – that is, one that has been carefully conceptualized and elaborated in accordance with the 80/20 rule. However, having said this, I would venture to observe that even the most refined or cutting-edge statistical procedures cannot succeed in "mending" a test whose content, elicitation procedures, and other design elements do not clearly and closely match up with the test's claimed or intended measurement purpose. In this regard, GIGA (or "garbage in, garbage analyzed") is really none too strong a term to characterize the meager and potentially misleading informational yield of testing undertakings that, for whatever reason, short-change the crucial initial planning and development steps and then look to statistical analyses to provide a miraculous cure for the test's inherent deficiencies.

Third and finally, I would suggest that we, as individual members of the testing community, have a tendency to be rather insular in our particular approaches to test development and associated research undertakings. I sincerely believe that we will enjoy much more progress and success as a profession if we intentionally and assiduously undertake to develop an operational ethic, together with expanded information and resource-sharing mechanisms, that will promote *more highly collaborative* use of the intellectual, institutional, and organizational resources available to us on a profession-wide basis, as opposed to, for the most part, simply "doing our own thing" in relative isolation as individual testers and researchers.

In this regard, ILTA's recently instituted web-based newsletter might provide a very appropriate mechanism for ILTA-wide discussion of the numerous ways in which an expanded role for the newsletter itself – supplemented by other electronic dissemination approaches as needed – could help move the language testing and research profession away from a largely "cottage industry" operation toward a much more consolidated enterprise – an approach that would provide greater benefit not only to ourselves as testers/researchers, but also to the many test-user communities we undertake to serve.

With thanks again to everyone, my very best regards, and every good wish for the future,

JOHN CLARK

BAAL/IRAAL 2006, Cork, Ireland

This year's annual conference of the British Association of Applied Linguistics (BAAL) was held jointly in collaboration with the Irish Association for Applied Linguistics (IRAAL) and was hosted by University College Cork in early September 2006. Fiona Barker and Lynda Taylor presented a paper entitled 'Using corpus studies to explore the impact of regional varieties of English on learners' written performance'. They described the use of the Cambridge Learner Corpus (CLC) and other corpus resources to search for evidence of possible regional variation in the written responses of FCE, CAE and CPE test takers.

Other papers presented a wide range of topics, many corpus-informed and relating to language teaching and assessment, including designing user-friendly corpus search facilities for non-linguists; creating a Business Word List; strategies used by distance language learners; primary/secondary language learning in the UK context and abroad, and beliefs about native-like English and World Englishes. Language assessment was discussed as a policy instrument and a number of papers explored aspects of the oral proficiency interview, such as staying on task and the influence of personality, gender and proficiency in group tests. Again, there were several thought-provoking papers which covered more unusual topics such as the position and role of Tongan people's oral practices in their experience of teacher training and attitudes towards 'inner-circle' regional varieties of English and other languages (in Guernsey, Channel Islands and Cork, Ireland).

ALTE language testing courses, Perugia, Italy

In September 2006, two week-long courses on language testing where organised by the Association of Language Testers in Europe (ALTE). These took place in Perugia, Italy and were hosted by ALTE member Università per Stranieri di Perugia. Attendees came from all over Europe and from a variety of testing organisations, including ALTE members (e.g. Goethe-Institut, Danish University of Education, Cambridge ESOL), ALTE observers (e.g. Scottish Qualifications Agency) and other non-ALTE members (e.g. Estonian Ministry of Defence).

The first week was an introductory course, led by Professor Cyril Weir (University of Bedfordshire) and Dr Barry O'Sullivan (University of Roehampton), which focused on the practical application of testing and assessment theory. Topics covered included: test design and test specifications, test production, quantitative and qualitative methods of test validation, and testing at different CEFR levels.

The second course focused solely on testing reading, again from a practical, evidence centred perspective. This was led by Professor Cyril Weir with assistance from Dr Hanan Khalifa (Cambridge ESOL). During the week topics for discussion included: a framework for testing reading comprehension, levels of proficiency, and cognitive, context, scoring, criterion, and consequential validity. Attendees were encouraged to apply testing theories to their own reading assessments and to share practices and ideas. The course ended with a practical session on item writing and preparing item writer guidelines, giving participants a chance to put the theories discussed into practice.

Further information about upcoming ALTE events can be found on the ALTE website: www.alte.org/further_info/index.php