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 21 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 developing materials for our language tests, taking into account the central test development issues of validity, reliability, impact and practicality (VRIP). It is also important for test developers to understand their test takers' experiential characteristics so that they can create tests that are appropriate to the target candidature.

In the opening article Lynda Taylor outlines the range of qualitative research carried out at Cambridge ESOL to support test development and monitor test quality. The following two articles describe question paper production (QPP) and the importance of candidate factors influencing performance. Tony Green and David Jay describe how question paper materials are pretested and reviewed and they present the eight stages of the QPP process which ensures quality assurance and quality control of all of our language tests. Next Hanan Khalifa considers whether test taker characteristics are accounted for in the Reading papers of the Cambridge Main Suite. She suggests how candidates' responses to Reading tasks may be affected by their physical/physiological, psychological and experiential characteristics, all of which form part of Cyril Weir's Validity framework.

Staying with Weir's socio-cognitive model, Stuart Shaw and Cyril Weir report ongoing research to articulate a clear theoretical and practical position for the construct of Writing which is an important component of all of our language tests. Weir's Validity framework attempts to reconfigure validity as a unitary concept, and to show how its constituent parts interact with each other. It identifies the various types of validity evidence that need to be collected at each stage in the test development process and identifies criterial parameters for distinguishing between adjacent proficiency levels.

Next Andrew Blackhurst reports on the latest trial of the computer-based IELTS test, looking at candidates' familiarity with computers and how examiner attitudes affected marking of writing scripts. Statistical analysis revealed no significant inter-group differences by gender, age or first language which suggests that the relationship between CB and PB scores is not affected by these differences between candidates.

We then review several recent publications including the latest Studies in Language Testing (SiLT) volume, *Testing the Spoken English of Young Norwegians*, which considers how communicative language ability (CLA) might be operationalised in the evaluation of the Norwegian speaking test for lower secondary students. Other recent publications are a review of the Cambridge Young Learners English Tests, an article on EAP study and score gains on the Academic IELTS module and a *Key Concepts* piece on washback and impact. Next we include two award announcements.

We end this issue with conference reports from two IATEFL events on learning English through picture books and new approaches to materials development, followed by an extended report on the second ALTE conference in Berlin, which Cambridge ESOL organised and contributed to in May.

We look forward to the Language Testing Forum happening in Cambridge in November.

Using qualitative research methods in test development and validation

LYNDA TAYLOR, RESEARCH AND VALIDATION GROUP

Introduction

Cambridge ESOL employs a wide range of research methods to support the process of developing materials for our English language tests and monitoring their quality. Over the last five years *Research Notes* has reported regularly on various studies which have adopted a quantitative or qualitative approach, including studies involving 'mixed methods'. The 'mixed methods' approach is one which is increasingly common among researchers today since it can enable us to gain a richer perspective in our research investigations. Such an approach is sometimes described as 'triangulation' – a term which refers to the use of multiple investigators, multiple theories or multiple methods, as well as to the use of different types or sources of data in order to cross-check the validity of findings (Trappes-Lomax 2004).

Given Cambridge ESOL's longstanding commitment to the direct assessment of speaking and writing ability, it's perhaps not surprising that qualitative research methods have come to occupy a central role in our research and validation programme in relation to these test components. The nature of the performance data produced from a speaking/writing test, combined with the complex interaction which constitutes direct performance assessment (see Milanovic and Saville 1996 for a discussion of the 'facets' of performance assessment), mean that qualitative research methods are often the best way to gain a richer, deeper understanding of the discourse and behaviours involved in speaking/writing assessment beyond the level of test score outcomes. Nevertheless, we have found qualitative research methods to be useful in investigating approaches to testing other skills including: indirect, task-based tests of reading/listening comprehension; form-focused tests of grammar, vocabulary and other language systems; and our more recent computer-based tests. This article gives a brief overview of key qualitative research methods used by Cambridge ESOL, the test contexts within which we use them, and the types of insights they can bring to the process of test development and validation.

The nature of qualitative research methods

Features of qualitative research are sometimes contrasted with those of quantitative research methods. Larsen-Freeman and Long (1991), for example, identified qualitative research as being: *naturalistic; observational; subjective; descriptive; process-oriented; valid; holistic; 'real', 'rich', 'deep' data; ungeneralizable; single case analysis*. Traditionally, the language testing community relied heavily upon the quantitative paradigm for its test development and validation activity, perhaps regarding a qualitative approach as too 'subjective' or not 'generalisable'; qualitative methods perhaps suffered from a 'Cinderella' status. In the mid 1980s, however, both Cohen (1984) and Grotjahn (1986) advocated using introspective techniques to better understand the testing process, rather than relying purely on the traditional statistical (i.e. quantitative) analyses; despite their recommendations, qualitative approaches to language test validation have only really began to impact on the field of language testing over the past 10–15 years.

Some of the research studies conducted by Cambridge ESOL (formerly UCLES EFL) in the early 1990s were among the first to apply the methodologies of discourse analysis and verbal protocol analysis in the language testing context. Professor Anne Lazaraton, currently at the University of Minnesota, is a pioneer of qualitative research in language testing and her involvement dates back to the late eighties when such approaches were not yet widely used in the field. It is in part due to her efforts that researchers are now more willing to embrace approaches that can provide access to the rich and deep data of qualitative research.

Using discourse/conversation analysis

Professor Lazaraton worked closely with Cambridge ESOL staff during the early 1990s in the area of oral proficiency assessment using qualitative discourse analytic techniques – particularly conversation analysis; the aim was to gain a deeper understanding of the speaking test event so that we could continually improve the quality of our speaking assessment. Between 1990 and 1992 work was conducted on the Cambridge Assessment of Spoken English (CASE) – an experimental speaking test developed largely as a research vehicle. This early work subsequently contributed significantly to the development of monitoring procedures for a wide range of Cambridge speaking tests. The work on CASE was followed by further conversation analytic studies of the Certificate in Advanced English (CAE) at CEFR C1 level and the Key English Test (KET) at CEFR A2 level; the focus was on interlocutor speech behaviour in both tests and on comparison across the two levels. This work contributed to the development of the 'interlocutor frame' – now a standard feature of Cambridge ESOL's speaking tests – as well as the development of examiner training and monitoring procedures. Qualitative analysis of candidate (as opposed to interlocutor) behaviour in the CAE, First Certificate in English (FCE) at CEFR B2 level, and International English Language Testing System (IELTS) speaking tests fed into the development of revised assessment criteria and rating scales. In addition, studies of FCE output explored the relationship between the task features in the four parts of the speaking test, and similar studies of IELTS in 1997 informed the subsequent revision of the Speaking Module

introduced in 2001. Studies of this type are ongoing and help us greatly in the development of new speaking tests and the revision of existing tests in relation to format, content, examiner training and the procedures necessary to monitor and evaluate how oral assessments are carried out.

The application of qualitative research methods to test-takers' writing performance has been equally fruitful since the early 1990s. In 1994, Cambridge ESOL started work on the Common Scale for Writing project with the aim of producing a scale of descriptors of writing proficiency levels. Phase 1 of this project involved a close, qualitative linguistic and functional analysis of a representative corpus of candidate scripts from PET, FCE, CAE and CPE examinations in order to generate 'can do', 'can sometimes do' and 'cannot do' statements. Previous reports in Research Notes, together with a recently published paper (Hawkey and Barker 2004) describe in more detail later work on this project which has included insights from computer corpus analyses as well as further qualitative textual analyses.

Using verbal protocol analysis

Discourse and conversation analytic techniques clearly lend themselves to analysis of performance data such as test-taker talk in speaking tests or essays from writing tests; they are less relevant, however, when investigating reading and listening tests where there is no 'performance artefact', such as an audio/video-recording or a written essay, which can be scrutinised during or after the event. In reading and listening performance – both the process and the product – is almost entirely internal and invisible to an observer. Assessment of reading and listening comprehension ability must therefore rely upon 'indirect' tests which seek to make visible the latent trait. Nevertheless, a different qualitative methodology – verbal protocol analysis (VPA) – offers one possible way of 'making visible' at least something of what may be going on inside a test-taker's head when they complete a reading or listening task. Alison Green's 1998 volume in our Studies in Language Testing series – entitled *Verbal protocol analysis in language testing research: a handbook* – offers an excellent introduction to this methodology.

One of the earliest studies to use VPA with a Cambridge ESOL test task was conducted by Taylor (1991). This was a small-scale exercise in which concurrent and retrospective think-aloud procedures were used with pairs to explore completion of an open cloze task from an FCE Use of English paper; the objective was to gain insights into the lexical, morphological, syntactic and semantic processing demands it made and the way these were sequenced. VPA was used again some years later (Whitfield 1998) to gain insights into test-takers' processing in the CAE reading comprehension test. The project arose out of the need to investigate difficulties, perceived and reported, encountered by CAE candidates in carrying out the gapped task (Part Two) on the Reading paper. The research exercise took place immediately prior to a live CAE test and the data collection instruments included: a test-taker's background questionnaire; a 'think-aloud' questionnaire; a 'talk-aloud' questionnaire; and a set of 'text-based' questions.

The use of VPA has provided us with rich insights not only into test-taker processing but also into the attitudes and behaviour of writing examiners. Studies in the early 1990s examined the marking strategies of examiners for FCE and CPE composition papers (Milanovic, Saville and Shen 1996). Further studies have taken place since then to explore rater attitudes and behaviour in the context of other Cambridge ESOL writing tests. The recently completed project to revise the IELTS assessment criteria and rating scales involved a comprehensive series of studies using VPA and other qualitative techniques (e.g. focus group discussion) to explore the attitudes and behaviour of IELTS examiners using the old and revised scales (see previous articles in *Research Notes*). Although VPA has been more widely used in the context of writing assessment, it has played a part in the development/revision of speaking assessment criteria and rating scales, and in the alignment of speaking tests at comparable levels across different suites. More recently, we have started to use VPA techniques with interlocutors and assessors in the Cambridge speaking tests to explore what they 'pay attention to' during a candidate's performance and how they arrive at a judgement about the quality of that performance.

Using observational checklists

Both the above qualitative methodologies – discourse/conversation analysis and VPA – can be challenging and time-consuming to work with, especially in relation to speaking assessment. In practice, discourse and conversation analysis require a great deal of time and expertise, and it can be impractical to try and deal with more than a small number of tests in any one study – thus making the results difficult to generalise. At Cambridge ESOL, therefore, we decided some time ago to explore an alternative and additional methodology based on using an instrument which could allow us to evaluate test-taker output in real time. The 'observational checklist' is based on a framework which describes the language of performance in a way that can be readily accessed by evaluators who are familiar with the tests being observed. It also has the potential for us to establish the relationship of predicted outcome to actual outcome using a data-set which satisfactorily reflects the typical test-taking population. The development and application of the observational checklist is described in articles in *Research Notes* 2, 3 and 11.

Other areas of research benefiting from qualitative approaches

Constraints of space make it impossible in this article to survey all the areas in which qualitative research methods support the process of developing and validating Cambridge ESOL's tests. However, three remaining areas of work are worth mentioning since these reflect current high-priority areas in assessment research not just for Cambridge ESOL but for the wider language testing community.

The first is the contribution which qualitative research can make in investigating aspects of test impact. Such methods can often allow the 'voices' of test stakeholders such as test-takers, teachers, administrators and score users to be heard more clearly. Recent research studies on the impact of IELTS (Green 2004 and Hawkey

forthcoming) and of PET in the Progetto Lingue 2000 (Hawkey forthcoming) show how qualitative and quantitative approaches can successfully combine to provide insights into the consequential validity of tests.

A second area, linked in some ways to the first, is the role that qualitative methods can play in benchmarking and standard-setting exercises, often when set alongside more quantitative studies. Over the past 18 months Cambridge ESOL has taken part in a series of empirical standard-setting projects with a wide range of key stakeholder users of our tests – the UK General Medical Council, the US National Council of States Boards of Nursing, the United Nations, and the Canadian Immigration authorities. All these projects have involved the use of panels of expert informants and judges – especially in relation to the speaking and writing subtests; the contribution of such panels is essential to inform an evaluation of various validity aspects of the test under scrutiny – theory-based, context, criterion-related and consequential.

Finally, qualitative method have an increasingly important role to play in improving our understanding of the test-taking processes on computer-based (CB) language tests, and especially how these may differ from the processes activated by conventional paper and pencil (PB) tests. The same is true for the processes adopted by examiners rating test performance on CB and PB tests. Language testing colleagues at CRTEC in Roehampton University have been working on our behalf over recent months specifically to investigate candidates' writing and reading test-taking processes on PB and CB tests. Elsewhere in this issue, members of the Research and Validation team discuss the importance of increasing our understanding of test-taking processes, including those relating to learning styles and strategies, interest/motivation, and special needs, and the relevance this has for any claims we might wish to make about the theory-based validity of our assessments. Qualitative research methods will undoubtedly continue to be central in our efforts to find out more about test taker characteristics at the individual and group level.

References and further reading

Cohen, A D (1984) On taking language tests: what the students report, *Language Testing* 1/1, 70–81.

—(1998) Towards enhancing verbal reports as a source of insights on test-taking strategies, in Huhta A, Kohonen, V and Luoma, S (Eds) *Current developments and alternatives in language assessment: Proceedings of LTRC 1996*, Jyvaskala: University of Jyvaskala and University of Tampere.

Green, A (1998) *Verbal protocol analysis in language testing research: a handbook*, Studies in Language Testing, Vol 5, Cambridge: UCLES/Cambridge University Press.

—(2003) Test impact and EAP: A Comparative Study in backwash between IELTS preparation and university presessional courses, unpublished PhD dissertation, University of Surrey, Roehampton UK.

Grotjahn, R (1986) Test validation and cognitive psychology: some methodological considerations, *Language Testing* 3, 159–185

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

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

Larsen-Freeman, D and Long, M H (1991) *An introduction to second language acquisition research*, London: Longman.

Lazaraton, A (2002) *A qualitative approach to the validation of oral language tests*, Studies in Language Testing, Vol 5, Cambridge: UCLES/Cambridge University Press.

—(2004) Qualitative research methods in language test development and validation, in Milanovic, M and Weir, C (Eds) *European language testing in a global context: Proceedings of the ALTE Barcelona Conference*, Studies in Language Testing, Vol 18, Cambridge: UCLES/Cambridge University Press.

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

Milanovic, M, Saville, N and Shen, S (1996) A study of the decision-making behaviour of composition markers, in Milanovic, M and Saville, N (Eds) *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium (LTRC), Cambridge and Arnhem*, Studies in Language Testing, Volume 3, Cambridge: UCLES/Cambridge University Press.

Taylor, L B (1991) *Some aspects of the comparability problem for communicative proficiency tests*, unpublished MPhil dissertation, Research Centre for English and Applied Linguistics, University of Cambridge, UK.

Taylor, L B and Saville, N (2001) *The role of qualitative methods in setting and maintaining standards in language assessment*, paper presented at the Setting Standards for Qualitative Research in Applied Linguistics Colloquium, American Association of Applied Linguistics Annual Conference, St. Louis, MO.

Trappes-Lomax, H (2004) Discourse Analysis, in Davies, A and Elder, C (Eds) *The Handbook of Applied Linguistics*, Oxford: Blackwell Publishing.

Whitfield, A (1998) *CAE Reading Research Project*, internal UCLES report.

Quality Assurance and Quality Control: Reviewing and pretesting examination material at Cambridge ESOL

TONY GREEN, RESEARCH AND VALIDATION GROUP
DAVID JAY, PRETESTING UNIT

Introduction

Cambridge ESOL is recognised by universities, employers and national education authorities around the world for its professional, rigorous and high standard assessments. This article will describe the vital role played by the reviewing and pretesting of material in ensuring the quality of our examinations.

Quality Assurance and Quality Control

When considering the effectiveness of an organisation in delivering quality, a distinction is often made between *quality assurance* and *quality control*. Quality assurance is concerned with processes and involves the management of activities and resources to improve benefits to stakeholders. Quality control is concerned with outcomes and involves checking that products meet intended standards. When producing examinations we follow procedures calculated to generate material of high quality, judge this material against established standards for quality control and feed back results to refine our processes.

Figure 1: The Question Paper Production process at Cambridge ESOL

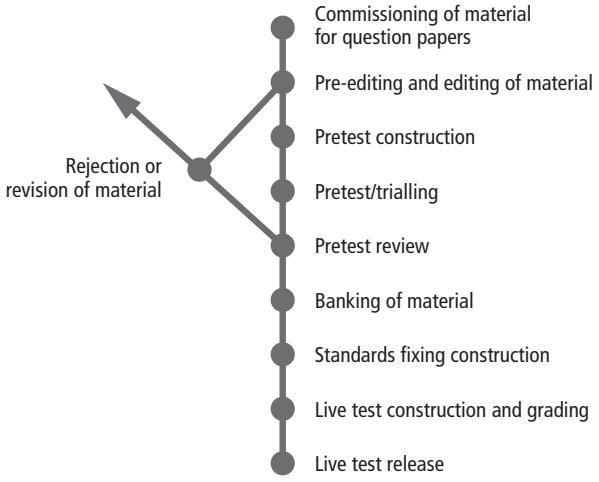


Figure 1 shows in outline the process of question paper production at Cambridge ESOL. At each stage in this process, from the initial commissioning of test material to the assembly of live test papers, there are checks in place to ensure that material reaching the live tests is of the highest quality. Quality control checks in question paper production take two forms: inspection and testing. Inspection involves people looking at new material to check whether it conforms to specifications; testing involves investigating the material to find out how it performs in use. Key occasions for inspection come at the *Pre-editing* and *Editing* stage and at *Test Construction*. Testing is of particular relevance to the *Pretest Review* and *Test Construction* stages.

Quality Control 1: initial inspection of the test material

Pre-editing is the first stage of the editing process and takes place when commissioned materials are initially submitted by item writers. A meeting is held involving the chairs of the item writer teams (experienced item writers) and Cambridge ESOL staff to consider the material. At this stage, guidance is given to item writers on revising items and altering texts, and feedback is provided on rejected texts and/or unsuitable item types. Routine checks at this point are intended:

- to ensure that all test material is culturally appropriate and accessible world-wide
- to ensure that all test material meets the test specifications
- to suggest appropriate changes to materials requiring amendments or rewriting.

With respect to these considerations, the *Pre-editing* process includes attention to the following features of the material:

- Topic
- Topicality
- Level of language
- Suitability for the task
- Length
- Focus of text
- Style of writing
- Focus of task
- Level of task

At *Editing*, texts and selected items are scrutinised again and are either approved for pretesting, amended, or, occasionally, may be sent back to a writer for further revision. Revised material will then be re-submitted for editing at a subsequent meeting.

Quality Control 2: testing the test material

In addition to the rigorous inspection of new material against test specifications, Cambridge ESOL also investigates the material in use before banking it for test construction. All material destined for use in live tests is first piloted through *Pretesting* (of papers containing items that can be marked objectively by computer or by clerical markers such as Reading papers) or *Trialling* (of papers that are scored by professionally trained examiners such as Speaking papers). New tasks are assembled into test forms and distributed to

a sample of candidates preparing for Cambridge ESOL examinations. Candidates participating in pretesting and trialling are broadly representative of the population taking the equivalent live test. Results from Pretesting are analysed by the Research and Validation Group and decisions are then made about whether the material can be banked for use in live tests. The following section describes some of the challenges presented by this process and the procedures in place to manage these effectively.

Quality Assurance: managing the pretest process

To commence the pretesting process, invitations are sent out to centres world-wide, targeting both registered and potential candidates for specific examinations. *Pretest windows* – the periods when pretests are made available to centres – are generally scheduled between 6 to 12 weeks prior to live test dates so that candidates for pretests will be taking the live examination soon afterwards, but will have enough time to benefit from feedback.

Response validity is enhanced by psychological and linguistic readiness on the part of candidates. Inevitably, there are substantive differences between pretest and live administration, not least the fact that the majority of pretests are invigilated by teachers in classrooms to students with varying degrees of commitment to taking a practice examination. Centres are therefore requested to place an emphasis on simulating the live experience in a 'dry run' of the examination day. It is important that the invigilation of the pretest should mirror that of the live examination. This level of authenticity is important, both for Cambridge ESOL and for the candidates themselves.

In the first instance, Cambridge ESOL relies on the good will and the commitment of centres to the quality of our examinations to encourage them to participate in pretesting. A significant source of motivation is that papers are returned to Cambridge ESOL for marking, with raw scores being returned to candidates within 3 weeks of receipt. However, as question paper security is paramount, there are limitations on the amount of feedback we are able to provide to candidates.

Finding enough pretesting candidates for our examinations presents Cambridge ESOL with a significant ongoing challenge, particularly for examinations with a more limited live candidature. To give some idea of the scale of this operation, almost 1,000 different pretest and trial papers were completed during the year 2003–4 with over 1,400 being scheduled for the year 2004–5. In order to meet the needs of examinations with limited candidature, in addition to the existing pretest windows, centres may be encouraged to increase candidature by the offer of pretesting on demand outside the allotted windows, through weekend 'Open Days' at larger centres or through a variety of incentive schemes.

For every live examination constructed, several pretest versions must be processed in order to accommodate the rejection or revision of test items. The average period between a pretest session and the appearance of the material on a live examination paper is 2 years. As a further guarantee of security, each live test paper is made up of tasks drawn from a number of different pretest versions.

The target specification for statistical analysis is a representative sample of 250 candidates per pretest version. In order to reach these numbers, over 300 question papers are allocated for each version and despatched world-wide. A further sample specification requires that there should be no more than 30% of speakers of any first language in a data set. All participants are volunteers and requests for numbers of papers sent to Cambridge ESOL by centres can only be estimates, so it is inevitable that rates of return are sometimes lower than expected. On average, just over 60% of papers are sent back completed. The returned sample may then be further pared down once item responses have been marked and optically scanned. Files passed on for statistical analysis will contain only those candidates who have completed the pretest correctly, have been entered at the appropriate level, and match the targeted balance of first language groups. Thus a proportion of completed pretests may be removed prior to statistical analysis. If the number of returned pretests is too low to meet the requirements material is re-allocated for use in the following pretest window.

In addition to the pretests, all candidates are also administered an anchor test and a background questionnaire. The anchor tests are made up of items of known difficulty, and this allows us to estimate the difficulty of each pretest version in relation to the established Cambridge ESOL Common Scale. Candidate background data (including, among other variables, age, gender and first language) enables Cambridge ESOL staff to further investigate, and if necessary, modify the sample for each pretest. Version-specific questionnaires for invigilators and candidates are also included and provide an invaluable source of qualitative feedback in evaluating the test material from the users' perspective.

Analysis of pretest results

At this point data files that satisfy requirements for representativeness are forwarded to Research and Validation for analysis. Objectively marked papers (papers that are marked clerically or by computer) are treated differently to subjectively marked papers (papers marked by qualified language teachers trained as examiners).

Objectively marked papers: Listening, Reading and Use of English

All candidate responses are analysed to ascertain the measurement characteristics of the material and match these against established standards. Both classical item statistics and latent trait (Rasch) models are used in order to evaluate the effectiveness of the test material[1]. Classical item statistics are used to identify the performance of a particular pretest in terms of the facility and discrimination of the items in relation to the sample that was used. Rasch analysis based on the anchor test is used to locate items on the Cambridge ESOL Common Scale of difficulty. In addition, comments on the material by the pretest centres and the immediate response of the pretest candidates are collected and taken into account in the selection process.

Subjectively marked papers: Writing and Speaking

Writing pretest scripts are marked by senior examiners and their comments are scrutinised to assess the suitability of tasks for

1. For more on the Classical and Rasch approaches see Bachman 2004.

inclusion in live test versions. The feedback on the trialling of the Speaking tasks is also assessed before material can be banked.

At a *Pretest Review* meeting, the statistics, feedback from candidates and invigilators and any additional information are reviewed in relation to Cambridge ESOL standards and informed decisions are made on whether texts and items can be accepted for construction into potential live versions. Material is then banked to await test construction.

Banking of material

Cambridge ESOL has developed its own item banking software for managing the development of new live tests. Each section or task is banked with statistical information as well as comprehensive content description. This information is used to ensure that the tests that are constructed have the required content coverage and the appropriate level of difficulty.

At regular *Test Construction* meetings, test papers are compiled according to established principles. Factors taken into account include:

- the difficulty of versions and the spread of items across defined ranges
- the balance of topic and genre
- the balance of gender and accent in the Listening versions
- the range of skills tested.

Our item banking software allows the test constructor to model various scenarios in order to determine which tasks should be combined to create tests that best meet the specifications.

Conclusion

As Cambridge ESOL continues to expand its range of examinations, generating, reviewing and pretesting sufficient material becomes ever more challenging. To safeguard quality we must continue to develop our quality assurance procedures and find new and more efficient means of inspecting and testing our material while enhancing their rigour. For the future, computer and internet-based solutions may provide promising new avenues for pretesting. Meanwhile, we continue to rely on the partnership with our test centres and will keep on looking for ways to add to the service we offer them through, for example, improved feedback and prompt turnaround of results.

Reference

Bachman, L F (2004) *Statistical Analyses for Language Assessment*, Cambridge: Cambridge University Press.

Are test taker characteristics accounted for in Main Suite Reading papers?

HANAN KHALIFA, RESEARCH AND VALIDATION GROUP

Introduction

This article examines test taker features that can affect the validity of responses to tasks and what an examination board like Cambridge ESOL can do to remove construct irrelevant barriers to test performance while maintaining the integrity of the construct being measured by an exam.

Weir (2005) argues that one of the key issues test developers are obliged to address is how the physical/physiological, psychological and experiential characteristics of candidates are catered for by a test. This article provides an account of how these characteristics have been addressed in the Reading papers of the Main Suite Examinations, that is, Certificate of Proficiency in English (CPE), Certificate in Advanced English (CAE), First Certificate in English (FCE), Preliminary English Test (PET), and Key English Test (KET).

Physical/physiological characteristics

Over the last decade increased attention has been given to issues of fairness in testing and test use, to the rights of test takers, and to testing candidates with disabilities (see Standards for Educational and Psychological Testing set by AERA, APA and NCME, 1999) with the practice of test accommodation usually perceived as promoting fair practice. Despite the controversy surrounding the use of test accommodations in terms of score interpretation and fairness to candidates with and without a disability, an extensive literature review of the effect of test accommodation on test performance by Sireci *et al.* (*forthcoming*) concludes that many accommodations are justified and are effective for reducing the construct-irrelevant barriers to test performance. A fairly consistent finding was that the accommodation of extended time improved the performance of students with disabilities. The review demonstrated a consensus on the fact that minor changes in the text associated with test items should not change the construct being measured. Abedi (2001:106) stated that *"modifying test questions to reduce unnecessary language complexity should be a priority in the development and improvement of all large scale assessment programs"*. The following description provides an account of how Cambridge ESOL Main Suite Exams deal with the physical/physiological aspects of test takers.

As an examination board, Cambridge ESOL is committed to ensuring that, as far as practicable, candidates are not excluded from taking an exam because of a disability or a learning difficulty, whether temporary or permanent. The principle behind this is that everything possible should be done to facilitate the candidates demonstrating their attainment in the skill being assessed. Special Arrangements are intended to remove as far as possible the effects of the disability on the candidate's ability to demonstrate his or her true level of attainment in relation to the assessment objectives; to ensure that candidates with disabilities are not given an unfair advantage over other candidates; and to avoid misleading the user of the certificate about the candidate's attainment. The most common provisions[1] for candidates with special needs, i.e. learning or visual difficulties, who are sitting for a reading paper are:

- **Additional time and/or supervised breaks** – candidates may require extra time to read their papers and write their answers. Examples of difficulties for which extra time might be appropriate include dyslexia or visual difficulties. For some candidates, supervised breaks may be appropriate instead of, or in addition to, the extra time allowance. An example would be a candidate who had difficulty concentrating for long periods of time.
- **Modified question papers** – candidates may require modified papers if they have severe visual difficulties. These may be in the form of contracted versions of Braille papers (where a single symbol may represent a group of letters) and un-contracted Braille (where there is a separate Braille symbol for every letter) or enlarged print.
- **Use of magnifying glass or hand-held scanning apparatus** – for partially sighted candidates.
- **Use of a Reader** – where a candidate has not yet learned to read Braille, Cambridge ESOL authorise a Reader to read out the whole of the examination. In these cases, the Reader is issued with a code of practice and if the reading texts are also read aloud, an endorsement is issued[2].

In modifying question papers Cambridge ESOL may seek the help of recognised organisations such as the National Institute for the Blind (RNIB) in the United Kingdom. These adapted versions cover the same assessment objectives as their standard counterparts with minor changes to rubrics, layout and sometimes length. They are produced to the same rigorous standards.

Endorsements are added to certificates where some of the objectives of the relevant examination have not been assessed on account of a particular disability of the candidate and where the candidate's performance in the examination was assessed on the basis of modified criteria to take account of particular learning disabilities, such as dyslexia.

Table 1 shows the number of provisions made for candidates taking Main Suite reading papers over a five-year period (2000–2004).

1. A comprehensive list of these provisions can be found on the Support page of www.CambridgeESOL.org under Special Arrangements.
2. It should be noted that in addition to the Reader, an invigilator must be present while the test is being administered.

Table 1: Special Arrangements provisions for candidates taking Main Suite papers 2000–2004

	2000	2001	2002	2003	2004
Total provisions for all papers	298	1045	1492	1494	1329
Provisions for Reading papers					
Additional Time and/or Supervised Break	140	392	544	700	587
Braille	14	33	33	28	31
Enlarged Print	34	58	66	68	65
Reader	N/A	1	1	5	9

Psychological and experiential characteristics

Another factor that should be taken into account as far as test taker features are concerned is the effect of affective and meta-cognitive domains on test performance. In other words how a test taker's attitude, beliefs, perceptions, self-esteem, interest, motivation, or anxiety may affect their performance on a given task.

O'Sullivan (2000) defines experiential features as those that are related to the test taker's educational and cultural background, experience in preparing and taking exams, as well as knowledge of the demands of a particular exam. Understanding the test taker's experiential characteristics would help test developers in creating tests that are appropriate to the targeted candidature.

The standards for educational and psychological testing set by AERA, APA and NCME (1999) advocate that test developers should provide the information and supporting evidence that test users need to select appropriate tests, and that test users should select tests that meet the intended purpose and that are appropriate for the intended test takers. The standards emphasise the joint responsibility of test developers and users to inform test takers about the nature of the test, test taker rights and responsibilities, the appropriate use of scores and procedures for resolving challenges to scores.

Research program

For several years, Cambridge ESOL has been working with researchers in the United Kingdom and North America to develop research instruments that will enable the development of a better understanding of the effects of the psychological, social and cultural contexts in which assessment takes place. Saville (2000) reports on one aspect of that research program which is the development of a bank of language learning questionnaires which investigates the background characteristics of ESOL candidature in relation to learning strategies and styles. The background factors are grouped as strategic, i.e., cognitive, meta-cognitive and communication strategies, and socio-psychological, i.e., attitudes, anxiety, motivation and effort. The questionnaires are intended to be used alongside examinations and tests in order to examine the relationships between test taker strategies and styles and their performance on language tests and on self-assessment instruments.

Exam related materials and activities

Cambridge ESOL aims to provide sufficient information on its examinations for decision-making on the part of test takers as well

as test users. Such information is available on its website, in exam-related handbooks, sample papers, teacher seminars, and examination reports. Part of the impact that Cambridge ESOL examinations have on the ESOL market is the availability of a number of coursebooks and practice materials by a range of publishers. Although Cambridge ESOL does not undertake to advise on textbooks or courses of study, it makes available on its website a list of publishers who produce materials related to the examinations.

Role of test taker characteristics within a validity framework

When considering test taker characteristics, Cambridge ESOL is naturally involved with other aspects of its approach towards building an evidence based validity argument for its examinations. ESOL's approach is one that acknowledges the importance of the socio-cognitive elements of validity and includes validity data that are theory-based validity, context-based, criterion-related, scoring-related validity and consequentially related validity as key elements.

Theory-based validity is concerned with the cognitive processing involved in carrying out a particular language task and the extent to which this represents the types of processing such activities would generate in real life. This is often referred to as interactional authenticity in the testing literature (Bachman and Palmer 1996). Context validity relates to the social arena in which an activity is performed i.e. the performance conditions under which a communicative activity takes place. It is concerned with the situational authenticity of a task (ibid). Scoring validity addresses the consistency and dependability of test results. Criterion related validity demonstrates how the test compares with other measures of the same construct either taken at the same time or at a later date. Lastly consequential validity examines the effects and impact of using the test or test scores.

There are obvious links between test taker characteristics and both theory-based validity and context-based validity in that individual characteristics will directly impact on the way individuals process the test task in terms of a particular configuration of contextual features. In fact in the very development of the test task the target test takers will have been taken into account in making decisions on such context parameters as topic, discourse mode and writer-reader relationships as well as the degree to which normal processing (theory-based validity) is exhibited by the target candidature. For example, LMS candidates differ from UMS candidates in terms of age group and purpose for taking an exam (see Table 2 for an overview of Main Suite candidature). Thus a task on the topic of hobbies and leisure may seem to be more appropriate to LMS candidates than a task on lifestyles and living conditions.

Test taker characteristics are taken into account at the pretesting stage for sampling purposes. Similarly, after implementation, when we are considering test results, we need to check that no group bias has been inadvertently introduced into the test in respect of any of the test taker characteristics which may affect consequential validity. Cambridge ESOL ensures that this is the case via, for example, the use of Candidate Information Sheets (CIS) which record information on test takers' age, gender, first language and so on.

Table 2: Overview of Main Suite candidature[1]

Exam	Candidature
KET	• Majority of candidates in Europe, South America & Asia–Pacific regions • 75% aged 18 or under, 50% aged 14 or under • 60% females • 85% attend preparation classes • 55% take KET out of personal interest, 40% for employment reasons, 30% are interested in further study of English.
PET	• Majority of candidates in Europe & South America regions • 70% aged 20 or under • 60% females • 85% attend preparation classes • 55% take PET out of personal interest, 50% to improve future employment prospects.
FCE	• Majority of candidates in Europe & South America • 75% aged 25 or under • 60% females • 80% attend preparation classes • 51% take FCE to gain employment, 32% for further study, 17% out of personal interest.
CAE	• Majority of candidates in Europe & South America • 80% aged 25 or under • 70% females • 80% attend preparation courses • 44% take CAE for study, 41% for work, 15% other
CPE	• Majority of candidates in Europe & South America • 75% aged 25 or under • 70% females • 85% attend preparation courses • Most candidates take CPE to work in their own country, then to work in another country and for further study of English and other subjects.

1. Information taken from: KET handbook (2003), PET handbook (2003), FCE handbook (2005), CAE handbook (2001 – to be updated Autumn 2005), CPE handbook (2002). All handbooks are available from www.CambridgeESOL.org

Conclusion

The care with which Cambridge ESOL considers test taker characteristics in designing test formats and materials continues to play a major role in test development and validation activities. Weir's Validity framework, described more fully in the following article, is providing an excellent means to consider our current practices and ways in which they could be improved.

References and further reading

Abedi, J (2001) *Language Accommodation for large scale assessment in science: Assessing English Language Learners*, (Project 2.4 Accommodation), National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

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

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

O'Sullivan, B (2000) *Towards a Model of Performance in Oral Language Testing*, unpublished Ph.D. dissertation, University of Reading.

Purpura, J E (1999) *Learner strategy use and performance on language tests: A structural equation modelling approach*, Studies in Language Testing, Vol 8, Cambridge: UCLES/Cambridge University Press.

Saville, N (2000) Developing Language Learning Questionnaires (LLQs), *Research Notes* 1, 6–7.

Sireci, S G, Scarpati, S, and Li, S (forthcoming) *Test Accommodations for Students With Disabilities: An Analysis of the Interaction Hypothesis*, Center for Educational Research Report No. 522, Amherst, MA: School of Education, University of Massachusetts Amherst. Revised version of paper to be published in AERA Review of Educational Research (personal communication with lead author).

Weir, C J (2005) *Language Testing and Validation: An Evidence-Based Approach*, Basingstoke: Palgrave Macmillan.

Establishing the Validity of Cambridge ESOL Writing Tests: towards the implementation of a socio-cognitive model for test validation

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

Introduction

This article reports briefly on the background to Research and Validation's work in articulating the Cambridge ESOL approach to assessment in the skill area of Writing. The perceived benefits of a clearly articulated theoretical and practical position for writing as a skill area underpinning Cambridge ESOL tests are essentially twofold. Within Cambridge ESOL we aim to deepen our understanding of the theoretical basis for how Cambridge ESOL tests different levels of language proficiency across its range of test products, and to inform current and future test development projects, especially in relation to Computer-Based Tests (CBT). Beyond Cambridge ESOL we hope to communicate in the public domain the theoretical basis for the tests and provide a more clearly understood rationale for the way in which Cambridge ESOL operationalise the theoretical constructs in its tests.

We are seeking to build on Cambridge ESOL's traditional approach to validating tests namely the VRIP approach where the concern is with Validity, Reliability, Impact and Practicality. As part of this renewal process the Research and Validation Group are presently exploring how far the socio-cognitive validity framework described in Weir's *Language Testing and Validation: an evidence-based approach* (2005) might contribute to an enhanced validation framework for our own thinking and activity. Weir's approach covers much of the same ground as VRIP but it attempts to reconfigure validity as a unitary concept, and to show how its constituent parts interact with each other. In addition it conceptualises the validation process in a temporal frame thereby identifying the various types of validity evidence that need to be collected at each stage in the test development process. Within each constituent part of the framework criterial individual parameters for distinguishing between adjacent proficiency levels are also identified.

Since December 2003 several members of the Research and Validation Group have been working directly with Cyril Weir to examine three skill areas comprehensively: Writing, Speaking, and Reading. Attempts to document the nature of the underlying constructs will result in the longer term in public statements, principally through volumes in the Studies in Language Testing series.

Weir's Socio-cognitive Framework for Validating Tests is described below and an explanation of how it is being employed for the validation of Cambridge ESOL Main Suite Writing test tasks is given. Progress on the forthcoming SiLT (Studies in Language Testing) volume *Examining Writing: research and practice* (Shaw and Weir in progress) is also reported. Even in its current draft, the documentation is proving to be useful to a wide range of Cambridge ESOL test developers and others such as item writers and Chairs (chairs are principally concerned with the technical aspects of writing the examination materials and ensuring that the writers are fully equipped to produce material to the best of their ability).

ESOL Test Development and Validation

In our existing approach to validation four essential qualities of test or examination usefulness, collectively known by the acronym VRIP (Validity, Reliability, Impact and Practicality), have been identified as aspects of a test that need to be addressed in establishing fitness for purpose (see Weir and Milanovic 2003 chapter 2). Cambridge ESOL examinations are designed around these four essential qualities, their successful validation being dependent upon all the VRIP features being dealt with adequately and completely.

Before the development or revision of a Cambridge ESOL

examination can be undertaken, a VRIP-based checklist must be constructed and a prioritised list of validation projects agreed and implemented. The necessary information which enables such a checklist to be compiled is collected through a process of successive cycles of consultation and trialling. Transparent and specific validation plans in the form of VRIP checklists are now used to ensure that all aspects of VRIP are appropriately accounted for a particular test thus corroborating any claims made about the usefulness of the test. The gathering of evidence, in the form of data collection, constitutes a principal consideration in the model based approach and provides the evidence to support the 'validity argument'.

The Cambridge ESOL approach to test validation is, however, an evolving one following on from the seminal work of Messick at the onset of the 1980s. In a recent position paper, Saville (2004) argues that "in order to develop a 'Cambridge ESOL validity argument', our test development model needs to be underpinned by theories (related to the VRIP features), in order to combine the *test development process* with *necessary evidence*" (Saville 2004:2). Weir (2005) provides a theoretical socio-cognitive framework for an evidence-based validity approach which accommodates and strengthens the existing VRIP approach.

The focus in this article is on Weir's socio-cognitive model which is ostensibly concerned with specifying and inter-relating focus areas for the validation process rather than with how the validation case should be argued *per se*. We would emphasise that related approaches such as those advocated by Toulmin (1958), Kane (1992), Mislevy et al. (2000), and Bachman (2004) are also under serious consideration in the development of a comprehensive approach to validation and the reporting of such at Cambridge ESOL. Of particular interest in this future development of our institutional approach to validation are evidence-centred assessment design and 'interpretive argument' logic.

We now turn to how Cambridge ESOL's Writing tasks are validated.

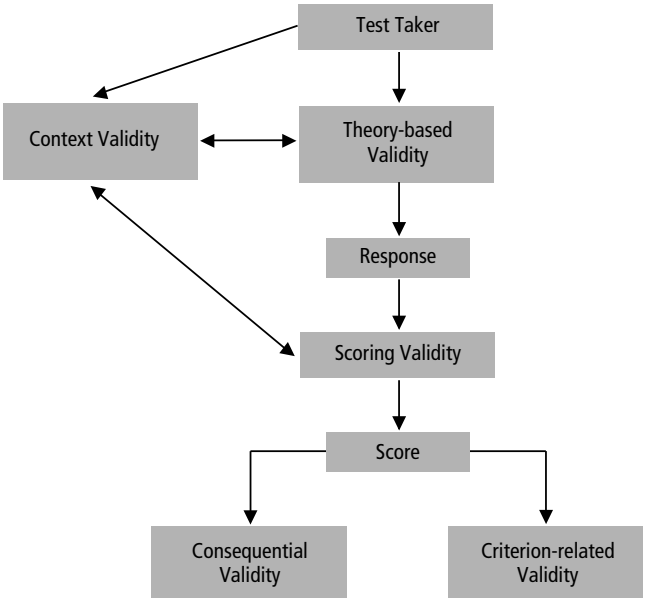
Cambridge ESOL Writing Tasks

The Main Suite general English examinations offer a comprehensive picture of how writing ability is measured by Cambridge ESOL across a broad language proficiency continuum. As such they constitute a major source of reference in the writing SiLT volume for illustrating how the writing construct differs from level to level in Cambridge ESOL examinations. In addition, frequent references to other writing papers from examinations in the Cambridge ESOL family such as IELTS, BEC, BULATS, and CELS are made to provide further clarification of how various performance parameters help establish distinctions between different levels of proficiency in writing and how research connected with these other examinations has had wider impact across all Cambridge examinations in relation to the provision of validity evidence for example in developing procedures to improve scoring validity. These non Main Suite examinations are well documented in their own right in other volumes in the SiLT series (see Hawkey 2004 for CELS, O'Sullivan forthcoming for BEC and BULATS and Davies forthcoming for IELTS) which offer a comprehensive coverage of their history, operationalisation and validity.

Validity of Cambridge ESOL Writing Tasks

The forthcoming SiLT volume on writing (Shaw and Weir) will offer a perspective on the central issues involved in the testing of writing in Cambridge ESOL examinations and will follow the conceptualisation of performance suggested by Weir (2005). A diagrammatic overview of the socio-cognitive framework is reproduced in Figure 1.

Figure 1: Weir's Validation framework



The framework is socio-cognitive in that the abilities to be tested are mental constructs which are latent and within the brain of the test taker (the cognitive dimension); and the use of language in performing tasks is viewed as a social rather than purely linguistic phenomenon. It represents a unified approach to establishing the overall validity of the test. The pictorial representation is intended to depict how the various validity components (and different types of validity evidence) fit together both temporally and conceptually. *'The arrows indicate the principal direction(s) of any hypothesized relationships: what has an effect on what, and the timeline runs from top to bottom: before the test is finalized, then administered and finally what happens after the test event'* (Weir 2005:43). Conceptualising validity in terms of temporal sequencing is of value as it offers a plan of what should be happening in relation to validation and when it should be happening.

The model comprises both *a priori* (before-the-test event) validation components of context and theory-based validity and *a posteriori* (after-the-test event) components of scoring validity, consequential validity and criterion-related validity. Weir comments thus on the complexity of the model:

> *'The more comprehensive the approach to validation, the more evidence collected on each of the components of this framework, the more secure we can be in our claims for the validity of a test.*

The higher the stakes of the test the stricter the demands we might make in respect of all of these.' (Weir 2005:47)

The *Test Taker* box connects directly to the theory-based and context validity boxes because *'these individual characteristics will directly impact on the way the individuals process the test task set up by the context validity box. Obviously, the tasks themselves will also be constructed with the overall test population and the target use situation clearly in mind as well as with concern for their theory-based validity'* (Weir 2005:51). Physical/physiological characteristics (individuals may have special needs that must be accommodated such as partial sightedness or dyslexia), Psychological characteristics (a test taker's interest or motivation may affect the way a task is managed or other factors such as preferred learning styles or personality type may have an influence on performance), and Experiential characteristics (the degree of a test taker's familiarity with a particular test may affect the way the task is managed) all have the potential to affect test performance (see Hanan Khalifa's article in this issue).

Context Validity

The term *content validity* was traditionally used to refer to the content coverage of the task. *Context validity* is preferred here as a more inclusive superordinate which signals the need to consider the discoursal, social and cultural contexts as well as the linguistic parameters under which the task is performed (its operations and conditions).

As a general principle it can be argued that language tests should place the same requirements on test takers as language does in non-test "real-life" situations. Bachman and Palmer (1996:23) describe a task as being relatively authentic '*…whose characteristics correspond to those of the Target Language Use (TLU) domain tasks'* and define authenticity as *'the degree of correspondence of the characteristics of a given language test task to the features of a TLU task'* (1996:23). Following Bachman and Palmer (1996), authenticity is considered to have two characteristics. Firstly, interactional authenticity (see section on theory-based validity below), which is a feature of the engagement of the test taker's cognitive capacities in performing the test, and secondly, situational authenticity (context validity in our terms) which attempts to take into account the situational requirements of candidates. Cambridge ESOL adopts an approach which recognises the importance of both situational and interactional authenticity.

Context validity in the case of writing tasks relates to the particular performance conditions under which the operations required for task fulfilment are performed (such as purpose of the task, time available, length, specified addressee, known marking criteria and the linguistic and discoursal demands inherent in the successful performance of the task) together with the actual examination conditions resulting from the administrative setting (Weir 2005:19).

Theory-based Validity

Theory-based validity involves collecting *a priori* evidence through piloting and trialling before the test event for example through verbal reports from test takers on the cognitive processing activated by the test task and *a posteriori* evidence involving statistical analysis of scores following test administration. This is necessary because language test constructors should be aware of the established theory relating to the language processing that underpins the variety of operations in real-life language use.

Scoring Validity and Criterion-Related Validity

Scoring Validity is linked directly to both context and theory-based validity and is employed as a superordinate term for all aspects of reliability. Scoring validity 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. Weir (2005:35) points out:

'For theory-based and context validity, knowing what the test is measuring is crucial. There is a further type of validity which we might term Criterion-Related Validity where knowing exactly what a test measures is not so crucial. This is predominantly quantitative and a posteriori concept, concerned with the extent to which test scores correlate with a suitable external criterion of performance (see Anastasia 1988:145, Messick 1989:16) with established properties'.

A test is said to have *Criterion-Related Validity* if a relationship can be demonstrated between test scores and some external criterion which is believed to be a measure of the same ability. Information on criterion-relatedness is also used in determining how well a test predicts future behaviour (ALTE 1998). Criterion-related validity naturally subdivides into two forms: concurrent and predictive. Concurrent validity seeks a 'criterion which we believe is also an indicator of the ability being tested' (Bachman 1990:248) and involves the comparison of the test scores with some other measure for the same candidates taken at roughly the same time as the test. This other measure may consist of scores from some other tests, or candidates' self-assessments of their language abilities, or ratings of the candidate by teachers, subject specialists, or other informants (Alderson et al 1995). Predictive validity entails the comparison of test scores with some other measure for the same candidates taken some time after the test has been given (Alderson et al 1995).

Consequential Validity

Messick (1989:18) argues that 'For a fully unified view of validity, it must … be recognised that the appropriateness, meaningfulness, and usefulness of score based inferences depend as well on the social consequences of the testing. Therefore social values and social consequences cannot be ignored in considerations of validity'. *Consequential Validity* relates to the way in which the implementation of a test can affect the interpretability of test scores; the practical consequences of the introduction of a test (McNamara 2000). Shohamy (1993:37) argues that 'Testers must begin to examine the consequences of the tests they develop … often … they do not find it necessary to observe the actual use of the test.'

Weir (2005) provides a comprehensive treatment of these key elements within the validation framework.

Relationships between different parts of Weir's Validity framework

Although for descriptive purposes the various elements of the model are presented as being independent of each other, there is a 'symbiotic' relationship between context validity, theory-based validity and scoring validity, which together constitute what is frequently referred to as construct validity. 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 completion. 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). Additionally criterion-related validity represents evidence of the value or worth of a test, and both will impact on the test (in terms of design, tasks etc.) and on the test taker. The interactions between and especially within these different aspects of validity may well eventually offer us further insights into more closely defining different levels of task difficulty. However, given our current limited knowledge of these effects, the separability of the various aspects of validity is maintained as they offer the reader a descriptive route through the model and, more importantly, a clear and systematic perspective on the literature.

Weir (2005:48) argues that test developers are obliged to seek to address all of the following questions:

- How are the physical/physiological, psychological and experiential characteristics of candidates catered for by this test? *(Test taker)*
- Are the characteristics of the test task(s) and its administration fair to the candidates who are taking them? *(Context validity)*
- Are the cognitive processes required to complete the tasks appropriate? *(Theory-based validity)*
- How far can we depend on the scores on the test? *(Scoring validity)*
- What effects does the test have on its various stakeholders? *(Consequential validity)*
- What external evidence is there outside of the test scores themselves that it is doing a good job? *(Criterion-related validity)*

Relevance for Cambridge ESOL

The documentation in its current state is already providing valuable insights to a range of key Cambridge ESOL personnel helping to:

- describe Cambridge ESOL's approach in skills assessment across levels and how this sits with theoretical context
- train Cambridge ESOL staff in skills assessment
- provide rationale/guidance on specific issues, such as rubric design
- develop item writer guidelines, ensuring coherence and preventing drift over time
- inform internal ESOL working groups and identify areas in need of attention according to the framework
- tag items for computer-based tests in terms of features likely to distinguish between levels
- inform the languages ladder (Asset Languages) project (see *Research Notes* 19)
- link exams to the CEFR
- inform content of ALTE Handbook and guidelines.

Conclusion

The issue of what a particular level of language ability means is critical for all aspects of language learning. Exam boards and other institutions offering high stakes tests need to demonstrate and share how they are seeking to meet the demands of context, theory-based, scoring, and consequential validity. In relation to these they need to be explicit as to how they in fact operationalise criterial distinctions between levels in their tests in terms of various parameters related to these. *Examining Writing: research and practice* marks the first attempt by any examination board to do this. Future research needs to investigate whether the parameters discussed in this volume either singly or in configuration can help better ground the distinctions in proficiency represented by levels in Cambridge ESOL examinations.

References and further reading

Alderson, J C, Clapham, C and Wall, D (1995) *Language Test Construction and Evaluation*, Cambridge: Cambridge University Press.

ALTE (1994) Code of Practice.

Anastasi, A (1988) *Psychological Testing* (6th edition), New York: Macmillan.

Bachman, L F (1990) *Fundamental Considerations in Language Testing*, Oxford: Oxford University Press.

—(2004) *Building and supporting a case for test ulilization*, paper presented at LTRC, March 2004.

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

Davies, A (forthcoming) *Testing Proficiency in Academic English 1950–2005*, Studies in Language Testing, Cambridge: UCLES/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, Vol 7, Cambridge: UCLES/Cambridge University Press.

Hawkey, R (2004) *A Modular Approach to Testing English Language Skills: The development of the Certificates in English Language Skills (CELS) examinations*, Studies in Language Testing, Vol 16, Cambridge: UCLES/Cambridge University Press.

Hawkey, R and Barker, F (2004) Developing a common scale for the

assessment of writing, *Assessing Writing*, 9/2, 122–159.

Kane, M T (1992) An argument-based approach to validity, *Psychological Bulletin*, 112/3, 527–535

McNamara, T F (2000) *Language Testing*, Oxford: Oxford University Press.

Messick, S (1980) Validity, in Linn, R L *Educational Measurement* (3rd ed), New York: Macmillan Publishing Company, pp.13–103.

Mislevy R J, Steinberg, L S and Almond, R G (2002) Design and analysis in task-based language assessment, *Language Testing* 19/4, 477–496.

Oller, J W (1979) *Language Tests at School: A Pragmatic Approach*, London: Longman.

O'Sullivan, B (forthcoming) *Issues in testing Business English: The revision of the Cambridge Business English Certificates*, Studies in Language Testing, Cambridge: Cambridge University Local Examinations Syndicate and Cambridge University Press.

Saville, N (2002) The process of test development and revision within UCLES EFL, in Weir, C J and Milanovic, M, 57–120.

—(2004) The ESOL Test Development and Validation Strategy, internal discussion paper, Cambridge ESOL.

Shohamy, E (1993) *The Power of Tests. The Impact of Language Tests on Teaching and Learning*, Washington, DC: NFLC Occasional Papers.

Toulmin, S E (1958) *The uses of argument*, Cambridge: Cambridge University Press.

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

Weir, C J (1988) Construct validity, in Hughes, A, Ported, D and Weir, C (Eds) *ELT Validation Project: Proceeding of a Conference Held to Consider the ELTS Validation Project Report*, The British Council and University of Cambridge Local Examination Syndicate.

—(1993) *Understanding and developing language tests*, New Jersey: Prentice Hall.

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

Weir, C J and Milanovic, M (Eds) (2003) *Continuity and Innovation: The History of the CPE 1913–2002*, Studies in Language Testing, Vol 15, Cambridge: UCLES/Cambridge University Press.

Listening, Reading and Writing on computer-based and paper-based versions of IELTS

ANDREW BLACKHURST, RESEARCH AND VALIDATION GROUP

Introduction

As reported by Tony Green and Louise Maycock in *Research Notes* 18 and 20, the Research and Validation Group has conducted a series of studies, since 2001, into the comparability of IELTS tests delivered on paper and by computer. Initial trials of the computer-based linear version of IELTS were encouraging, finding that test format had little effect on the order of item difficulty and finding strong correlations between scores on the CB and PB versions of Listening and Reading forms, suggesting that format had a minimal effect on the scores awarded. However, one potential limitation of the original trial design was that candidates knew that the computer test was not for real. Consequently further trials were undertaken in 2003, in which the candidates took both paper-based and computer-based versions of the exam, not knowing which (or indeed whether both) would generate their actual results. Green and Maycock reported on the results of the first phase of these trials, Trial A: this article looks at the results of Trial B, and at the continuing validation work involved in the roll-out of computer based IELTS to a number of test venues this year.

Other comparability studies

Cambridge ESOL's research into computer-based tests has taken place in parallel with work undertaken elsewhere. Choi, Sung Kim and Boo (2003), for example, reported on an examination of the comparability between paper-based and computer-based test versions of the Test of English Proficiency developed by Seoul National University. Their findings supported the hypothesis that the PB version and the CB versions of the TEPS subtests (listening comprehension, grammar, vocabulary, and reading comprehension) did indeed produce comparable results.

One particular focus of interest for researchers has been whether candidates' familiarity with computers might have an impact on their exam performance. During the CB IELTS trials, as reported in *Research Notes* 20, candidates were asked to complete a questionnaire specifically to procure information about their experience of, and confidence in using, computers.

With the great expansion in the use of computers in the office, at school and in the home over the past fifteen years, it is interesting to note that the nature of researchers' interest in candidates' familiarity with computers may be changing. When Bunderson et al (1989) offered an overview of studies of test equivalence, they commented: "In general it was found more frequently that the mean scores were not equivalent than that they were equivalent; that is the scores on tests administered on paper were more often higher than on computer-administered tests" (p378). While they held those differences to be, in general, quite small and of little practical significance, their concern was that lack of familiarity with computers might be a factor in producing lower scores on computer-based tests.

By contrast, Russell and Haney (1997) looked at the test

performance of school students in Worcester, Massachusetts. Here the problem was that scores appeared low, and the researchers' experiments tested their theory that since most student assignments were completed using computers, but the tests were in paper-and-pencil mode, this change in format was adversely impacting student writing scores. While conceding that their study was quite small and might not be generalisable, they concluded that "estimates of student writing abilities based on responses written by hand may be substantial underestimates of their abilities to write when using a computer" (p. 16).

Another issue has been differences in the way that examiners approach typed and handwritten scripts. As Bennett (2003) observed, available research tends to suggest that typed essays receive lower scores, possibly because substantive and mechanical errors stand out more, since the responses are easier to read. Brown (2000) investigated differences between handwritten and word-processed versions of the same IELTS Task Two essays. Handwritten versions of the same script tended to be awarded higher scores than the word-processed versions, with examiners apparently compensating for poor handwriting when making their judgements. Shaw (2003) in *Research Notes* 11 reported similar findings for First Certificate scripts.

On the other hand, a study by Whitehead (2003) reported in *Research Notes* 10 found no significant differences between scores awarded to handwritten and typed IELTS scripts. Although CB scripts had yielded slightly lower scores and higher variance, Whitehead suggested that these differences could be attributable to a motivation effect given that the candidates knew that they would not actually be assessed on the computer test.

The results of trial B

A total of 785 candidates took part in Trial B. As in Trial A, candidates were allowed to choose whether to answer the written test by hand or on computer. For the purpose of analysis, a sample of 467 candidates was constructed such that:

- only candidates for whom a complete data set of PB and CB results was available were included, numbering 622 candidates
- no single first language group would constitute more than 30% of the sample: a random, stratified sample was taken from the 622 'valid' candidates to achieve this
- the sample was broadly representative of the IELTS population with regard to other factors such as gender, reasons for taking the examination, levels of education completed and age ranges.

Due to the self-selective nature of the trial, some groups (for example, the younger age ranges) were slightly over-represented in the sample. However, a broad range of candidates were represented.

The mean band scores for the forty-eight paper-based Reading versions administered between September 2003 and August 2004 fluctuated between 5.58 and 6.35, and those for Listening (for Academic candidates) fluctuated between 5.79 and 6.48. Since the mean band scores for the computer-based components administered during Trial B fall within the range of band scores obtained during the period in which the trial was conducted, and we have no reason to suspect that the ability of the Trial B sample differs significantly from that of the live population (given the band scores obtained), we might conclude that CBIELTS is grading candidates at approximately the correct level.

The correlations given in Table 1 are sufficiently large to suggest that the rank ordering of candidates does not differ significantly between different modes of administration. The values may appear to be lower than one might expect, but note that there is variation in the reliability of the different test forms and skills, which cannot be estimated accurately (since the paper-based scores are taken from a number of test versions). Thus, the difference in reliability can not be corrected for by the calculation of disattenuated correlations.

Table 1: Correlations between scores on different skills and in different modes

Skill		Reading		Listening		Writing	
	Mode	PB	CB	PB	CB	PB	CB
Reading	PB	1					
	CB	**0.712**	1				
Listening	PB	0.717	0.62	1			
	CB	0.725	0.692	**0.764**	1		
Writing	PB	0.625	0.554	0.66	0.669	1	
	CB	0.564	0.504	0.622	0.641	**0.658**	1

Cross tabulations of the paper-based and computer-based scores were constructed and the rates of agreement were calculated for each component (see Table 2). Half band scores used in reporting performance on the Reading and Listening components of IELTS typically represent two or three raw score points out of the 40 available for each test.

Table 2: Agreement rates after removal of outliers

	Reading	Listening	Writing[1]	Overall[2]
Kappa	0.1290	0.1596	0.2805	0.4006
% agreement	25.52%	25.06%	45.01%	49.88%
% agreement to within half a band	68.45%	66.13%	45.01%	95.59%
% agreement to within a whole band	90.95%	88.17%	83.76%	100%

1. Scores for Writing tests are awarded in whole band increments

2. Note that overall scores for the two tests (CB and PB) include a common Speaking component

These may be compared with the following agreement rates for live and preparatory candidates taking two paper-based test versions two weeks apart, reported by Thighe (2001) in Table 3.

Thus the rates of agreement between paper-based IELTS and CB IELTS are satisfactorily similar, when compared with the agreement rates for live candidates taking two paper-based IELTS versions.

Table 3: Agreement rates of live and preparatory candidates given in Thighe (2001)

	Live candidates		Preparatory candidates	
	Reading	Listening	Reading	Listening
% agreement	30%	27%	27%	25%
% agreement to within half a band	68%	62%	61%	68%
% agreement to within a whole band	89%	89%	85%	91%

The reliability of the CB versions of the test used in Trial B is indicated by Cronbach's alpha, which for Listening version 50001 was 0.893 and for academic Reading version 50001 was 0.816. Both of these values are within the range we would expect for ordinary live versions (for which the historical range is 0.710–0.897). The pattern of the distribution of band scores for the computer-based components is broadly in line with those for the paper-based components. The distributions for Writing are especially similar.

The analyses conducted into writing performance on the CB and PB versions of IELTS indicated, in common with Whitehead (2003), that there are no significant differences between scores obtained on the CB and PB versions of the Academic Writing test. Although there was some evidence that reflected Brown's (2000) concern that legibility may impact on rating, the actual impact on scores appeared minimal.

Are different groups of candidates affected differently by the format of the test?

The performance of males and females on paper-based IELTS and computer-based IELTS is compared in the table below. There is no evidence to suggest the existence of any gender bias although both genders scored slightly higher on CB Reading than PB Reading and lower on CB Listening than PB Listening. Females did equally well on CB and PB Writing whereas males did less well on the Writing paper when taken in the CB format.

Table 4: Band scores by gender

Gender		Reading	Listening	Writing	Speaking	Overall
Female	PB	5.92	6.17	5.81	6.14	6.09
	CB	6.26	5.85	5.81	-	5.95
Male	PB	6.03	6.18	5.73	6.14	6.08
	CB	6.35	5.86	5.66	-	5.94

The largest first language group represented in the trial was composed of Chinese speakers. The scores for Chinese and non-Chinese candidates in Trial B are compared in Table 5. For Reading and Listening, the average shift in score for Chinese candidates is not significantly different to the average shift in score for non-Chinese candidates. It is worth noting that, in Trial B, both groups secured on average a lower score in the CB Listening test. In Trial A, both Chinese and non-Chinese speaking candidates obtained on average a higher score in both skills. This is evidence that there is no systematic tendency for candidates to perform better on one mode than the other.

Results on the Writing test were further investigated, using repeated measures analyses of covariance (ANCOVA), with PB writing test scores as dependent variable, to explore differences between groups in the relationship between paper and computer-based scores. Groups were defined by gender (Male or Female); age (five different age groups) and first language (Chinese or Non-Chinese L1). Handwritten responses to the CB test were separated from word-processed responses for the purpose of the analysis.

ANCOVA revealed no significant ($p>0.01$) inter-group differences by gender, age or first language either where CB scripts had been typed or handwritten. This suggests that the relationship between CB and PB scores is not meaningfully affected by these differences between candidates. There were also no significant ($p>.01$) differences between scores on the CB and PB tests, when responding on paper or on screen, either for the Chinese L1 or Non-Chinese L1 groups. These results suggest that the CB and PB versions of the IELTS Writing test yielded comparable scores across groups.

Conclusion

The data gathered since 1999 has provided evidence that CB IELTS can be used interchangeably with PB IELTS, and that candidates, given adequate computer familiarity, will perform equally well on either version of the test. Accordingly, a limited number of IELTS centres are now offering CB IELTS as a live test. In reporting results, no distinction will be made between candidates who have taken the test in one mode or the other: from the point of view of receiving institutions the results may be considered equally valid, whichever form was taken.

Since the live trials were concluded, the revised Writing assessment criteria and scales have been introduced. Further studies will be undertaken to assess the impact of these changes on the marking of typewritten scripts, and we will also be seeking feedback from examiners involved in marking typewritten scripts from the live test. We will also be studying the reading and listening performance data generated in this initial phase of CB IELTS, as there is expected to be a different profile of first languages among the live candidates, as compared to that obtained in the trials. In a separate project, as mentioned in *Research Notes* 18, Cyril Weir, Barry O'Sullivan, and colleagues at the Centre for Research in Testing, Evaluation and Curriculum at Roehampton University, have been commissioned by Cambridge ESOL to investigate candidates' reading test taking processes on CB and PB tests. Their work will consider questions regarding the processes in which candidates engage, and the nature of the language elicited, when taking tests with different formats and will appear in a future issue.

There are important considerations involved in providing the test in different formats: not all candidates will have adequate computer familiarity; some candidates may experience fatigue when reading extended passages on computer. Accordingly, Wolfe

Table 5: Band scores by first language

Gender		Reading	Listening	Writing	Speaking	Overall
Chinese	PB	5.81	5.99	5.66	5.76	5.87
N=140	CB	6.20	5.59	5.49	-	5.70
	Difference	-0.38	0.39	0.17		0.18
	Correlation	0.682	0.676	0.559	-	0.873
Non-Chinese	PB	6.03	6.26	5.82	6.30	6.18
N=327	CB	6.35	5.96	5.86	-	6.05
	Difference	-0.31	0.29	-0.03		0.13
	Correlation	0.720	0.793	0.685	-	0.914

and Manalo (2004) recommended that test designers "think seriously about providing examinees with a choice of composition medium ... particularly when high-stakes decisions will be made based upon the test results" (p. 61). The IELTS partners have always recognised that it is important that candidates should be able to take the test in the form with which they feel comfortable: the pen and paper test will continue to be available, so only those candidates who feel confident in their ability to use a computer need do so, and, in the live test as in the trial, candidates taking the CB test will have the option of responding to the writing test by hand. In this way, as the new form of the test becomes more widely available, IELTS will ensure that candidates have the option of taking the test that suits them best.

References and further reading

Beeston, S (2000) The UCLES EFL Item Banking System, *Research Notes* 2, 8–10.

Bennett, R E (2003) Online Assessment and the Comparability of Score Meaning, *paper presented to International Association for Educational Assessment Annual conference*, Manchester, October 2003.

Brown, A (2000) Legibility and the rating of second language writing: an investigation of the rating of handwritten and word-processed IELTS Task Two essays, *IELTS Research Projects 1999/2000*.

Bunderson, C V, Inouye, D K and Olsen, J B (1989) The four generations of computerised educational measurement, in Linn, R L (Ed.) *Educational Measurement* (3rd ed.), American Council on Education, New York: Macmillan,.

Choi, I-C, Kim, K-S, and Boo, J (2003) Comparability of a paper-based language test and a computer-based language test, *Language Testing* 20/3, 295–320.

Green, A (2004) Comparison of Computer and Paper Based Versions of IELTS Writing: A further investigation of Trial A data, Cambridge ESOL Internal Validation Report 585.

– (2005) Composing and scoring CB scripts: Analysis of CB IELTS Trial A and B Writing data, Cambridge ESOL Internal Validation Report 643.

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

Maycock, L (2004a) CBIELTS: A Report on the Findings of Trial A (Live Trial 2003/04), Cambridge ESOL Internal Validation Report 558.

—(2004b) Candidate use of the ten-minute transfer time provided in the paper-based IELTS Listening component, Cambridge ESOL Internal Validation Report 584.

—(2004c) CB IELTS: A Report on the Findings of Trial B, Cambridge ESOL Internal Validation Report 605.

Popham, W J (1988) *Educational Evaluation (2nd ed)*, New Jersey: Prentice Hall.

Russell, M and Haney, W (1997) Testing writing on computers: An experiment comparing student performance on tests conducted via computers and via paper-and-pencil, *Educational Policy Analysis Archives*, 5/3.

Shaw, S (2003) Legibility and the rating of second language writing: the effect on examiners when assessing handwritten and word-processed scripts, *Research Notes* 11, 7–10.

Shaw, S, Jones, N and Flux, T (2001) CB IELTS – A comparison of computer based and paper versions, Cambridge ESOL Internal Validation Report 216.

Thighe, D (2001) IELTS PB and CB Equivalence: Comparison of Equated Versions of the Reading and Listening Components of the IELTS Paper Based Examinations, Cambridge ESOL Internal Validation Report 288.

Wolfe, E, and Manalo, J (2004) Composition medium comparability in a direct writing assessment of non-native English speakers, *Language Learning and Technology*, 8/1, 53–65.

Recent publications of interest

Recent months have seen the publication of several items which may be of interest to readers of *Research Notes*. A new volume has appeared in the *Studies in Language Testing* series and an edited version of the series editors' note for this is given below; three more SiLT volumes – including two reporting on case studies of washback/impact, and one on the assessment of business English – are planned for publication before the end of 2005. In addition, several well-known refereed journals have recently included contributions relating to Cambridge ESOL tests or written by Cambridge ESOL staff.

Studies in Language Testing – Volume 20

Volume 20 – entitled *Testing the Spoken English of Young Norwegians* – reports on a two-part study to validate a test of spoken English for Norwegian secondary school pupils (EVA). The study, undertaken by Angela Hasselgreen, involved a corpus-based investigation of the role played by 'smallwords' – such as *well, sort of, and you know* – in bringing about fluency.

Following an introduction in Chapter 1, Hasselgreen goes on to provide in Chapter 2 a clear exposition of the nature of test validation and offers a comprehensive working framework for the validation of a spoken language test. It is interesting to compare the extent to which Hasselgreen's broad conceptualisation of this area matches the operational procedures for test validation adopted by Cambridge ESOL in terms of Validity, Reliability, Impact and Practicality (VRIP) - as these are described in Volumes 15 and 16 of the Studies in Language Testing series. Together they provide a solid grounding for any future work in this area.

Chapter 3 examines in detail how communicative language ability (CLA), a central element of a test's theory-based validity, might be operationalised in the evaluation of the Norwegian speaking test for lower secondary school students of English (EVA). As such it represents one of the few reported attempts to operationalise Bachman's seminal cognitive model of language ability.

In Chapters 4 and 5 she takes the broader validation framework developed in Chapter 2 and applies it to the EVA test and so provides test developers with a working example of how validation might be done in practice. She was able to evaluate all aspects of communicative competence in EVA as it had been defined in the literature to date. Published studies of this type are regretfully rare in the testing literature and Hasselgreen's case study illuminates this vital area of our field in an accessible, well written account of a validation exercise carried out on this spoken language test in Norway.

Her validation of the existing test system throws up serious problems in the scoring instruments. In particular the band scale relating to fluency does not adequately account for the aspects of CLA measured by the test, particularly as regards textual and strategic ability, because it lacks explicit reference to the linguistic devices that contribute to fluency. Low inter-rater correlations on *message* and *fluency*, discussed in Chapter 5 in the discussion of *a posteriori* validation based on test scores, further point to the problem of vagueness in the existing definitions of these criteria. This provides the link to the second part of the monograph: how to establish 'more specific, unambiguous, data-informed ways of assessing fluency'. As such it addresses the emerging consensus that rating scale development should be data-driven.

In Part 2 of her study Hasselgreen focuses on one aspect of the validation framework that frequently generates much discussion in testing circles, namely how we should develop grounded criteria for assessing fluency in spoken language performance. In Chapter 6 she examines the relationship between smallwords – such as *really*, *I mean* and *oh* – and fluency at different levels of ability. According to Hasselgreen such smallwords are present with high frequency in the spoken language and help to keep our speech flowing although they do not necessarily impact on the content of the message itself. A major contribution of this monograph is the way she locates her argument in relevance theory as the most cohesive way of explaining how smallwords work as a system for affecting fluency by providing prototypical linguistic cues to help in the process of interpreting utterances.

In Chapter 7, based on a large corpus, she reports her research into the extent to which students taking the EVA test used smallwords. She used three groups of students: British native speaker schoolchildren of 14–15 years of age, and a more fluent and less fluent group of Norwegian schoolchildren of the same age allocated on the basis of global grades in the speaking test. The results support the case that the more smallwords a learner uses, the better their perceived fluency. Critically she found that the more fluent speakers of English clearly used this body of language more frequently than high and low achieving Norwegian learners, and the range of the words they used was larger especially in turn-internal position to keep them going. The more fluent learners used smallwords in a more native like way overall than the less fluent; they used them in most turn positions, and also with a greater variety of forms and uses. More native like quantities and distribution of smallwords 'appear to go hand in hand with more fluent speech'. The clear implication is that because smallwords make a significant contribution to fluent speech, such features have an obvious place when developing effective fluency scales. In Chapter 8 she analyses in more detail how students use their smallwords in helping create fluency in communication - what small words actually do, providing further corroboration of the findings in Chapter 7.

In Chapter 9 she looks at background variables in relation to smallword use such as gender and context, and considers the acquisition of smallwords. She then looks at the implications of the findings of her research for language education, assessment (task and criteria), and for teaching and learning. Chapter 10 summarises the data in relation to the original research questions.

This volume presents the reader with a valuable framework for thinking about test validation and offers a principled methodology for how one might go about developing criteria for assessing spoken language proficiency in a systematic, empirical manner.

Items in refereed journals

The Test Reviews section of the April 2005 issue of *Language Testing* (22/2) contained an independent review of Cambridge ESOL's Young Learners English (YLE) Tests by Alison Bailey of UCLA, USA. In Part I of her review, Bailey begins with a useful overview of key features of YLE such as test purpose, administration, scoring procedures, test length and price, etc. She goes on to describe in general terms the content of the Listening, Reading and Writing, and Speaking subtests. Part II of the review evaluates the strengths and weaknesses of the YLE tests in terms of the essential test qualities of validity, reliability, fairness (developmental appropriateness and cultural sensitivity), practicality (administration and scoring), and impact. Her concluding summary in Part III describes the Cambridge YLE tests as 'superior tools to

most other options for assessing young learners' on the grounds that 'they were developed for the EFL learner with their specific learning situation in mind'. At the same time, she helpfully highlights issues which can inform our current and future research agenda for the YLE tests. Cambridge ESOL welcomes this type of independent, professional evaluation of our tests and we were pleased to be able to respond to specific questions and requests for information during preparation of this particular review.

A paper by Tony Green was recently published in the journal *Assessing Writing* (10/1); entitled 'EAP study recommendations and score gains on the IELTS Academic Writing test', his paper reviews recent research relating to score gains on the IELTS test and reports on two linked studies of gains made on the academic writing module. Phase 1 involved over 15,000 candidates taking the official test on two occasions and Phase 2 involved nearly 500 learners on English for academic purposes (EAP) courses taking the IELTS Writing test at course entry and exit. The study's findings offer insights into the amount of intensive English study which learners will require if they are to improve their band scores, with initial scores proving to be a stronger predictor of outcomes than course length. In the face of changes over recent years in the international student population, insights from this and similar studies could be used to update study recommendations for students and academic institutions such as those published by the British Association of Lecturers in English for Academic Purposes (BALEAP).

Finally, the April 2005 issue of the *English Language Teaching Journal* (59/2) included a short contribution by Lynda Taylor on the topic of washback and impact in their Key Concepts in ELT section. This section of the journal aims to assist readers to develop an appreciation of central ideas in ELT, and to approach the content of articles from a perspective informed by current debate on aspects of theory and practice.

Award announcements

UCLES/ILTA Lifetime Achievement Award 2005

The UCLES/ILTA Lifetime Achievement Award 2005 has been awarded to Professor Bernard Dov Spolsky, of Bar-Ilan University. In an announcement to the ILTA discussion list, committee members Lyle Bachman, Vivien Berry, Carolyn Turner and Nick Saville described how Bernard's work in language testing over the past 40 years has provided seminal insights and observations that have stimulated the field in a number of areas.

Two of Bernard's early articles, "Language testing: the problem of validation" (1968) and "What does it mean to know a language, or how do you get someone to perform his competence" (1968) were among the first not only to question the nature of the ability that we intend to measure, but also to link validity to the uses we make of language tests. His empirical research in the 1960s and 70s challenged the conventional wisdom, based on structural linguistics, that language proficiency consisted essentially of bits and pieces of knowledge, and laid much of the conceptual groundwork for subsequent investigations in the field into the nature of language ability.

Bernard's long-standing and continued concern with ethical issues in language testing has also guided and stimulated the field. His *magnum opus* is, without a doubt, *Measured Words* (1995) in which he traces the history of language testing in Europe and the US, as this was played out in institutional language testing on both sides of the Atlantic.

In addition to his own research and writing, Bernard has served and guided the language testing community through his membership on many national and international professional committees and as President of the International Language Testing Association. The UCLES/ILTA Lifetime Achievement Award will be presented to Professor Spolsky at LTRC 2005 in Ottawa, Canada. For more information about LTRC see www.carleton.ca/ltrc/

ILTA Honorary Lifetime Membership

Dr Caroline Clapham, a former Cambridge ESOL colleague, has been awarded an honorary membership of ILTA by the Executive Board. This was awarded in recognition of her outstanding service to ILTA.

Caroline has made a long and dedicated commitment to ILTA and language testing in general. Caroline's contributions to the profession started in the 1970s with her work on the UK General Medical Council's test for doctors, the Professional and Linguistics Assessment Board (PLAB). She was also involved in the development of the original ELTS test, now the IELTS test. She wrote an excellent Ph.D, that raised and elaborated significant questions about the testing of English for specific purposes and was joint-winner of the TOEFL award for best dissertation in 1996; it is now a much referred to volume in the SiLT series. The Encyclopedia volume that she edited with David Corson (Kluwer) is also much cited. She was Editor of *Language Testing Update* with Dianne Wall for many years and was also a long serving member of the Editorial Advisory Board of *Language Testing*, offering reliable and constructive reviews for the journal.

For three decades Caroline has given her support and careful advice unstintingly to a new generation of language testers across the world and is widely revered by her former students as an amazing doctoral supervisor. Her last appointment was as IELTS Validation Officer for Cambridge ESOL until ill health sadly forced her to retire.

Conference reports

The last six months have been a busy time for Cambridge ESOL staff attending and organising conferences and events. In this section we have included a series of conference reports covering two IATEFL events, one on learning English through picture books (Trish Burrow) and another on new approaches to materials development for language learning (Andy Blackhurst), together with an extended review of the contributions of Cambridge ESOL staff to the recent ALTE conference which took place in Berlin in May (Andrew Balch, Margaret Cooze, Angela Wright and other Cambridge ESOL staff).

IATEFL YL SIG event: Learning English through Picture Books

An IATEFL Young Learners Special Interest Group (YL SIG) event was held at the International Youth Library, Schloss Blutenburg, Munich from 19–21 November 2004. This conference was organised by London Metropolitan University, Realbook News and the Mopsy Club in conjunction with the IATEFL YL SIG, and with support from the British Council and several publishers. The conference drew together over 160 teachers, teacher trainers, experts in YL methodology, authors and publishers from Germany, neighbouring European countries and countries as far afield as Australia and South Africa.

The initial idea for the conference was conceived at an informal meeting of the YL SIG at the IATEFL conference in Brighton in April 2003. Members felt that against the background of continued growth in teaching English to young learners (TEYL), the role of picture books in the classroom was still relatively unresearched. Despite the fact that a picture book-based methodology is not a recent development, it was felt that this area remained undiscovered by many teachers, or that they used picture books only as an occasional supplementary activity. This conference thus grew out of the desire to make stories and picture books more central to children's English language learning experience.

Janet Enever from London MET and Opal Dunn of Realbook News opened the event with a plenary entitled 'Why Picture books?' The presenters started by outlining the value of picture books in the YL classroom and gave examples of how they contribute to language acquisition and to children's broader educational development. They argued that through the use of appropriate visuals and texts, picture books can aid language learning. Several different types of picture books were cited as well as how they can help to develop the ability to respond to and use varied features of language.

Reference was made to research which shows that most EYL lessons involve published materials, mainly in the form of coursebooks. Whilst most TEYL materials contain a story, these are often written to fit the coursebook and highlight a specific feature of language, and so do not constitute a picture book. The speakers emphasized that they wished to advocate the use of picture books, not as a replacement to the coursebook, or merely as an occasional 'treat', but as an important part of lesson. The rationale given is that pictures repeat and confirm meaning and therefore serve to extend children's understanding of a text and stimulate their imaginations.

Several questions, which it was hoped the conference would answer, were raised in the plenary. They included:

- What is a text?
- Is the diet we are giving children too sparse?
- How can we make picture books more widely available?
- How do boys and girls respond to picture books?
- What is the role of visuals in supporting learning?
- What is the role of written text in children's future lives and how can we support learning appropriately?

The plenary concluded with a call for more research to be undertaken and for existing research to be disseminated more widely. The audience was left with these research questions to consider:

- How do YLs use picture books in constructing theories about other languages and cultures, themselves and their emotions and their relationship with others?
- What might picture books contribute to the process of developing oral skills?

In the second plenary, Gail Ellis, The British Council's Global Manager for Young Learners, and Carol Read, a freelance educational consultant, teacher trainer and writer gave an introduction to the British Council's Magic Pencil teaching materials[1]. The Magic Pencil exhibition and website celebrates children's book illustration and brings together the work of 13 illustrators. Both offer foreign language teachers of children and teenagers a valuable resource. The website makes a wide range of teaching materials based on picture books available to primary and secondary EYL teachers, and are linked to the Common European Framework levels A1 and A2. The presenters outlined aims of the Magic Pencil exhibition and materials, such as the desire to bridge the gap between assisted reading in class and enabling children to read independently.

On Saturday morning, participants were able to choose from four different research strands. In each strand a group of four teacher researchers gave 30-minute presentations describing their classroom-based research projects which were followed by a methodology session. At the end of the day, reporters who had been stationed in each of the four sessions gave a short summary of the morning's research projects and some findings.

In the afternoon sessions devoted to the topic of illustration, participants were able to choose two from three presentations given by authors, illustrators and publishers. Tony Ross, a writer and illustrator of children's books, including *Dozy Mare* and

1. These can be viewed at http://magicpencil.britishcouncil.org/

Misery Moo, spoke about how he creates stories. He gave a humorous and also thought-provoking account of how he writes stories that address children's hopes and concerns by devising stories for his own children and developing the ideas. He closed the session by giving a visual demonstration of an idea he is currently developing for a story called *Tadpole's Promise*.

Tessa Strickland, the co-founder and publisher of Barefoot Books, also writes children's stories under the pen name of Stella Blackstone. She described how the company started from a small, home-based beginning and is now an international publisher. Using one of her own books, from the *Cleo the Cat* series, she showed how illustration and rhyme can hold the interest of small children and help children with very limited language to understand and participate in stories. She also described how Barefoot Books encourages a cross-cultural approach by matching illustrators and writers from different cultures to produce appealing books which gain and sustain children's interest.

At the end of Saturday, in a session entitled 'Taking Stock', the four reporters gave summaries of the issues raised in the four research strands in order to give participants a snapshot of what had been discussed in the research sessions which they could not attend. Some of the highlights were: Olga Vrastilova's account of how Primary school teachers in Eastern Bohemia have worked to incorporate children's literature in their EYL classes, and how this has informed changes to pre-service training given to teachers. Penelope Robinson discussed the classroom practice of using Big Books to promote L1 and L2 development and argued that there was a need to develop a methodology which fully exploits the language learning potential of this practice. Sandie Mourão outlined her research into understanding and encouraging greater English storybook borrowing in Portuguese pre-schools.

In her research presentation, Carol Read argued that effective classroom talk and interaction is required when using picture books in order to scaffold children's learning. Silvana Rampone focused on how cross-curricular project work which involves using picture books and fact books linked to a specific topic can aid the development of L2 literacy. Anneta Sadowska's presentation argued that the introduction of a three-year reading programme had led to improved language use, evidenced by higher scores in the Cambridge Young Learners English Tests.

The four methodology sessions addressed varied areas of concern to many YL teachers. Opal Dunn gave an account of how boys and girls differ and the implications this has for teachers using picture books. Annie Hughes and Heide Niemann considered the importance of selecting the 'right' picture book to trigger children's imaginations and support learning of useful learning strategies. Sandie Mourão demonstrated how teachers can effectively use a picture book to help children learn to tell stories and progress to making a classroom book to share with friends and family. Gail Ellis outlined how a story-based methodology has evolved in the years between the first edition of *Tell It Again!*, a story-based methodology book that she wrote in 1991 with Jean Brewster and which has been recently revised to fit current TEYL practice.

The conference resumed on Sunday morning with an open session which gave participants the chance to demonstrate how they had used a favourite picture book with their class. The enthusiasm of the teachers and teacher trainers was evident in their animated presentation of the stories and follow-up activities they had devised. These included the creation of a class storybook which featured photos of the class on their journey around the school, travelling past the playground, over the hill and so on.

In the panel discussion, which was chaired by Prof. Dr Gisela Schmid-Schönbein, a range of questions which had been gathered from participants the day before were submitted to the panel for comment and discussion.

In the closing plenary, Dr Friederike Klippel, Chair of English Language Education at Ludwig-Maximilians-University, Munich, addressed the question of what literacy involves. She stated that it is easier to define what a picture book is than why picture books are important. She emphasized the importance of exploring the aesthetic of picture books and using them to fulfill basic educational aims of school, as opposed to seeing them only as a vehicle for presenting information or information about language. She referred to the constraints placed on language learning, which is often not given much time in the school timetable. This echoed the research presentations of Saturday, in which teachers described very varied teaching contexts where English is given from as little as 30 minutes a week to a full 4 hours. She raised the question of providing appropriate teacher training which gives teachers the confidence to use picture books, adding that the rapid growth in TEYL has meant that many non-native speaker teachers may not be trained to teach either English or the age group in question.

The crucial role that artwork plays in supporting learning was addressed as was the importance of helping children to create an inner picture, as this mental representation helps YLs to make sense of the story and to relate it to themselves. Dr Klippel raised the issue of the quality of discourse in classrooms and cited 1990's research in Vienna which found that less than 1% of all classroom talk comprised storytelling events. Dr Klippel closed her plenary by proposing a new way of conceiving literacy; one which sees aesthetic and affective aspects as integral to the meaning and teaching of a text.

In her round-up, Opal Dunn spoke about the need to address the development of emotional literacy in YL classrooms. She made a call for teachers to be creative so that they enable children to explore the rich world of picture books. She acknowledged the budget constraints of many schools, but urged teachers to establish ways in which all children can have access to the rich language learning experience constituted by picture books, for example through school fundraising events.

The issues and questions raised at this event have direct relevance to the Cambridge YLE Tests. In the test development phase, Cambridge ESOL recognised the value of a story-based methodology in the YL classroom, and this is reflected in the use of stories in tasks like the storytelling in the Movers and Flyers Speaking tests and in all three levels of the Reading/Writing papers. Visual literacy is central to the last task in Starters Reading/Writing, where children answer a series of questions about a story depicted in three pictures. Events such as this emphasize the value of using picture stories and also the need to conduct more research into how stories and picture books can support young learners in the learning of a second language.

BALEAP/SATEFL Conference: New Approaches to Materials Development for Language Learning

On 15th–17th April, the James Watt Conference Centre, at Heriot-Watt University, was the venue for a conference jointly organised by the British Association of Lecturers in English for Academic Purposes (BALEAP) and the Scottish Association for the Teaching of English as a Foreign Language (SATEFL). This first such collaboration between the two organisations proved very successful, with some 250 participants registering. The theme of the event was 'New Approaches to Materials Development for Language Learning'.

Several speakers touched on issues of inter-cultural communication, reflecting a concern to develop materials and tasks which would assist overseas university students to adjust to the western academic context, and placing emphasis on raising individual student awareness of subtle cultural differences in expectations about academic teaching and learning, and familiarising the newcomers, both linguistically and culturally, with their new learning environment, inside and outside the classroom, while at the same time sensitising the host community to their needs.

There was also much debate about the ownership of English: for example, what are the implications of the role of English as an international language, even as a lingua franca, for the teaching and assessment of the language? Linked to this was discussion of the challenges of adapting teaching materials produced for a global market to local contexts, and the impact on learners overseas, particularly young learners, when the language is contextualised in western settings, or when presented in more familiar settings.

Cambridge ESOL was well-represented at the conference. Besides a stand in the Conference Exhibition, two presentations were given during the parallel sessions. Lee Knapp's presentation 'I know, therefore I teach: evaluating knowledge for teaching', introduced the recently developed Teaching Knowledge Test (TKT), developed as part of an alternative framework of qualifications intended to cater more closely to the needs of non-L1 teachers of English, offering candidates a step in their professional development as a teacher.

Discussion of the impact of new technologies was a natural part of a conference on new materials. Students are increasingly familiar with using computers and accustomed to generating written work on computer. As described elsewhere in this issue, the IELTS partners have responded to this new environment by launching a linear computer-based version of the IELTS examination. Sharon Jordan and Andy Blackhurst's presentation at the conference described the development of the computer-based test, and reported on the research studies conducted into the comparability of the new format and the existing paper based form. Since May 2005, the CB version of IELTS has been available as an optional alternative to the current paper-based IELTS test at a number of centres.

This event incorporated a lively exchange of practical ideas and much to provoke further thought on materials development and the differing linguistic and cultural needs of language learners.

ALTE Second International Conference, Berlin

The Goethe-Institut hosted ALTE's (Association of Language Testers in Europe) Second International Conference in Berlin from 19–21 May 2005. The conference was entitled 'Language Assessment in a multilingual context – Attaining Standards, Sustaining Diversity' and set out to explore issues of ethics, quality and transparency.

The conference opened with a plenary delivered by Suzanne Romaine (University of Oxford). Romaine focused on a number of responses to recent charges of a growing lack of linguistic diversity across the world. While the influence of English continues to grow and tests of linguistic ability for citizenship in many countries are introduced, languages worldwide continue to die out. Romaine illustrated how language communities are becoming weaker and the family is losing its place as primary language agency, with the educational system often becoming the replacement, if one exists at all.

Several lesser-used languages were represented at the conference, some still used as first languages, and some being actively revitalised. Test differences can be illustrated by looking at the speaking components of two of the tests discussed at the conference. Emyr Davis (Welsh Joint Education Committee) outlined the development of an entry level qualification in Welsh for adult learners. Davis discussed the priorities and aims of this test and showed how it mapped onto the Common European Framework of Reference (CEFR) as well as highlighting one of the main intentions of the test which was to focus on spoken language. To this end the lower level tests of Welsh are weighted to carry 50% of available marks for the speaking component of the test. The corresponding level of test from the Cambridge ESOL Main Suite carries 25% of the total available mark.

Nele Maddens (Centre of Language and Migration, Leuven, Belgium) demonstrated the training involved in administering the speaking component of the Certificaat Nederlands als vreemde Taal – the Certificate of Dutch as a foreign language. While Cambridge ESOL speaking tests have a Team Leader system which operates worldwide, this test relies on volunteer Dutch teachers acting as interlocutors for the speaking test, which is recorded so that assessment can be carried out centrally in Leuven. Video support is available to guide and train interlocutors and to ensure standardisation.

Lynda Taylor (Cambridge ESOL) continued on the theme of linguistic diversity with her paper considering the implications that language varieties have for testing (see her summary below).

Day two began with a plenary by Elana Shohamy (University of Tel Aviv) arguing that language tests are used as covert policy tools in multilingual societies, giving out messages about language priorities and leading to exclusion, lack of representation and violation of language rights. She raised the question of whether it is necessary to know a language in order to be a good citizen, a question that was discussed throughout the day's Language Testing and Citizenship Forum. It emerged from the forum that there are considerable differences between the systems of the various countries presented (in terms of the language level demanded, the existence and extent of a 'citizenship test', the overall price of

tests, and whether tests had to be taken before arrival in the country in question). There was agreement that these were issues requiring further discussion, and that ALTE should be involved.

The morning plenary on the last day of the conference was given by Brian North (Eurocentres) and presented the Common Reference Levels of the CEFR, outlining the origins and aims of the CEFR and describing the Swiss project which set out to develop and scale the descriptors in the 1990s. Work on relating examinations to the framework is ongoing, and a number of sessions described projects that have set out to do this. Firstly, Waldemar Martyniuk (Council of Europe) discussed the piloting of a manual which sets out to define exam specifications in relation to the framework and to ensure a consistent interpretation of the levels. The final version of the manual is expected to be published in 2008. Then Henk Kuijper and Aukje Bergsma (Citogroep, Netherlands) described their research which set out to determine whether the State Examination for Dutch as a second language (used as a component of university entrance requirements) meets the needs of both candidates and universities, and also how the examination relates to CEFR levels. A needs analysis was conducted using a selection of CEFR 'can do' statements. Candidate performances were then assessed both in relation to these statements and according to the normal assessment criteria of the examination. It was assumed that the assessment according to the CEFR statements was reliable, and the number of candidates 'incorrectly' passing or failing the examination for any given pass mark was calculated. The final step was of particular interest to test developers, in that the researchers decided that, in the case of university entrance examinations, false negatives were worse than false positives, and should therefore be double-weighted. The ideal pass mark was then identified as that for which the total number of (weighted) false results was smallest. This was found to coincide with the current pass mark for most skills. The researchers felt that in contexts such as aviation, where safety takes priority, false positives should probably be considered more important than false negatives.

The contributions made by Cambridge ESOL staff at this event are summarised below.

Standardising speaking tests for the languages ladder: A case study across 3 languages and levels

Karen Ashton presented the methodology and results from a case study looking at issues of comparability in recently developed speaking tests in Spanish, French and German for Asset Languages, one of Cambridge ESOL's new products (see www.assetlanguages.org.uk for further information).

This case study covered three languages (French, German and Spanish) across three levels of the framework; Breakthrough, Preliminary and Intermediate. Issues of comparability across these three languages were discussed as Karen detailed the process used to create an observational checklist which was used by raters to rank order candidates and to create profiles of candidates at each level. Additionally Karen discussed provisional alignments of these assessments for Asset Languages to both the CEFR and to qualifications within the UK.

Asset Languages gained a lot of interest at the ALTE Berlin conference and there was an interesting discussion on relating speaking assessments to the CEFR following the presentation.

Test comparability and construct compatibility across languages

Peter Hardcastle (Cambridge ESOL), Sibylle Bolton (Goethe Institute) and Francesca Pelliccia (Università per Stranieri di Perugia) presented a paper on the ALTE European Language Test Validation Project. This has been working since February 2004 on establishing parameters of comparability among language tests in German, Italian, Spanish, Portuguese and English, specifically with a view to aligning language tests to the CEFR. The Goethe Institute, the University of Salamanca/Instituto Cervantes, the University of Lisbon, the Università per Stranieri di Perugia and Cambridge ESOL are the participating members in the scheme which has been examining the extent to which tests and task-types can be compared using linguistic criteria, performative or pragmatic criteria and cognitive criteria and how these dimensions can be used to establish cross-language test equivalence. The objective is to make a strong and empirically justifiable claim that a preliminary (say, CEFR A2) test of Italian measures proficiency in that language at the same level as a preliminary test in another European language. In other words, passing an A2 test of Italian means that the Italian proficiency level of that test taker is similar to the proficiency level of another test-taker who passed an A2-level test in Portuguese, Spanish, German or English. Such claims have been difficult to make convincingly, though efforts continue to be made through schemes such as the Council of Europe's European Language Test Benchmarking Project.

The presenters suggested that IRT (one-parameter Rasch) scaling is probably the most convincing way to align language tests to the CEFR at the moment, pending more thorough studies into the various linguistic and cognitive features of test tasks, in which test difficulty is thought to reside. Such scaling techniques are currently being applied to the Spanish and German tests in the project and will shortly be extended to include the Portuguese and Italian tests. The Goethe Institute and the University of Perugia demonstrated how the project had brought significant benefits to their testing programmes since the February 2004 inception.

Methods for constructing a multilingual framework: The Languages Ladder Project

Neil Jones gave a presentation on the Languages Ladder. This is a 'can do' description of a set of levels, and a key part of the National Languages Strategy – an initiative to address poor levels of foreign language competence in England. Asset Languages is the name of the assessment system being developed by Cambridge ESOL and OCR for the Languages Ladder. The project, which is to include at least 26 languages, both those taught as foreign languages as well as those used in communities, implies a complex framework which requires robust methods for cross-language equating. The presentation outlined a methodology for developing this framework for the objectively-marked skills of Reading and Listening, based on scale construction using IRT methods. It is critical that the procedure can be replicated across languages: if the overall framework is to be coherent then there is

very limited freedom to make judgements about how levels are assigned for particular languages.

A model was presented for defining a progression of levels of maximum use to learners, on the principle that each level should represent an achievable target and a substantive learning gain. This was compared with the Cambridge ESOL levels, which have developed organically over a long period of time in a way which appears, interestingly, to reflect these requirements.

A methodology was presented, based on the above model, for developing scales and equating them across languages, and then illustrated by work done so far on scale construction for the Asset languages currently under development (French, German, Spanish).

Standard-setting remains a critically important aspect of framework construction. The methods currently proposed, e.g. in the Council of Europe's Pilot Manual for equating language tests to the CEFR, place too great an emphasis on test-centred approaches. The interpretation of test performance concerns what learners can do in the world beyond the test, and so it is learner-centred standard-setting which offers the best chance of achieving genuine equivalence across languages.

Testing Teaching Knowledge: developing a quality instrument to support professional development

Through this presentation Hanan Khalifa Louhichi raised awareness of the quality assurance procedures followed during the planning, design and development phases of a new Cambridge ESOL product, namely TKT (Teaching Knowledge Test). The paper started with an overview of Cambridge ESOL's model-based approach towards test development followed by a discussion of the iterative and consultative nature of the test development process and the necessity to build in validation checkpoints at the outset of test construction.

The presenter then provided an account of how the teaching knowledge construct has been defined and operationalised in TKT before describing a series of research activities and quality assurance procedures that have been carried out as part of TKT development. These included findings from the trialling phase in Asia, Europe, and Latin America, from stakeholders' feedback questionnaires, and from standard-setting activities. A framework of building validity evidence was also briefly discussed. The paper concluded with an overview of on-going validation activities and projected future research studies.

Considering Young Learners: The Cambridge Young Learners English Tests

Helen Spillett and Juliet Wilson gave a talk on the Cambridge Young Learners English (YLE) Tests. They discussed the features of an effective and appropriate language test for children and considered how well the Cambridge YLE Tests exemplify these.

They looked at the model for test development and revision used within Cambridge ESOL and described how it has been applied in the current revision of the Cambridge YLE Tests. In particular they looked at the consultation process and the development and trialling of revised tasks. They described the way in which the quantitative and qualitative data were analysed before the final task specifications were drawn up. They also considered the importance of monitoring the impact of the revision and of continuing to review the Cambridge YLE Tests in the light of relevant research.

In the presentation, they considered some of the challenges Cambridge ESOL faces as a provider of tests taken by children all over the world. These include ensuring that the tests are fair to all candidates in terms of the construct of the tasks, the marking process and cultural accessibility. Participants were invited to look at a sample illustration to prompt discussion about its suitability for young learners from a range of cultural backgrounds.

Linguistic diversity: language varieties and their implications for testing and assessment

Lynda Taylor described how the worldwide spread of English over several centuries has led to the emergence of regionally-based varieties (e.g. American/British/Australian English); more recently 'new Englishes' have emerged in certain regions (e.g. Hong Kong/Singaporean/Euro-English). Other European languages (e.g. French, Spanish) have experienced a similar evolution, and can point to both well-established and emerging regional varieties, all with their distinctive features.

The phenomenon of language varieties raises some interesting issues for language test developers. Most language proficiency tests restrict themselves to a single language variety, usually a standard 'native speaker' (NS) variety with 'prestige' status. Recently, however, the usefulness of traditional NS language models has been questioned. Some linguists argue that the nature of international communication today requires that teaching and testing reflect multiple language varieties on pragmatic and equity grounds; others maintain that well-established standard NS varieties are necessary if teaching materials and tests are to have international currency. This debate touches upon the norms, standards and criteria we adopt for teaching/assessment, as well as the role played by 'non-native' and 'native' speaker teachers and assessors.

This paper considered the issues that linguistic diversity raises for those involved in language teaching and testing provision. Findings from a recent survey of perceptions, policy and practice among European test providers were presented and discussed. The presentation concluded with suggestions on how testing agencies can adopt a principled and pragmatic approach to this issue – one which acknowledges and affirms linguistic diversity while at the same time maintaining standards of quality and fairness.

The 2nd ALTE conference was a great success and highlighted some of the challenges faced by Cambridge ESOL and its ALTE partners both in setting standards for multilingual, multi-level projects such as Asset Languages (which will provide assessment in 26 languages at 16 levels), and for ensuring that tests meet the needs of the test-takers. Meanwhile, Lynda Taylor's call for a principled approach to policy relating to linguistic diversity has implications for all language tests and testers.

For further information on the work of ALTE visit their website at www.alte.org