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 22 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 ethics in testing, primarily as it relates to test fairness in general and more specific ways. These related concepts inform the development and revision of language tests and impact more widely on our approach to providing appropriate language tests and teaching awards. This issue includes articles on computer-based testing, assessing writing, rater behaviour and also describes how test fairness impacts on exams testing general, business and academic English, as well as exams testing children and adults.

In the opening article Nick Saville describes how the Association of Language Testers in Europe (ALTE) sets professional standards for its members through a Quality Management approach. This involves the adoption of a Code of Practice and associated systems and practices. For Cambridge ESOL this means a continual process of striving to maintain the quality of all of our products, whether they are general or domain-specific language tests, or our range of teaching awards.

Next Juliet Wilson discusses some of the ethical issues concerning testing children. She outlines Cambridge ESOL's current approach to testing this group of learners and describes how these tests were developed in the mid 1990s in terms of their design and the children's experience of taking these tests. Once such issues are identified in relation to a particular examination, Cambridge ESOL staff use their experience and current thinking in the field of language testing to operationalise such issues which then feed into our ongoing test development and revision cycle.

The following articles consider computer-based testing, a format which allows for more flexible and secure tests which have associated benefits for candidates, administrators and examination boards. Paul Seddon presents an overview of computer-based tests before Ed Hackett illustrates the development of CB PET, focusing on how paper-based materials have been adapted for computer-based testing to suit the candidature which enables equivalence with the paper-based format.

Staying with the paper-based versus computer-based theme, Stuart Shaw reviews the literature on word processed text and evaluates the impacts for the assessment of both writing quality and rater behaviour. He concludes that examiner training should ensure equity between the rating of these two formats.

We follow this with an update of two key areas of research and development: computer-based testing (CBT) and Asset Languages, both of which are concerned with equal access to language awards and providing fair and accurate assessments to candidates. Next we review this year's staff seminar programme and the latest Studies in Language Testing (SiLT) volume on test impact.

We end this issue with conference reports from events Cambridge ESOL staff attended during the summer months, including the presentation of two awards at the LTRC conference in Ottawa in July. Our conference activities will continue during the winter with our hosting of the Language Testing Forum in Cambridge in November.

Setting and monitoring professional standards: a QMS approach

NICK SAVILLE, RESEARCH AND VALIDATION GROUP

Introduction

This article provides an update on earlier information which I provided in *Research Notes 7* (February 2002) on the work of the ALTE Code of Practice Working Group. This time I provide a brief overview of the approach which has been adopted by ALTE and how it relates to other attempts to address the same issues. In particular, I focus on the most recent work of the working group which deals with the implementation of an auditing system within the ALTE Quality Management approach.

Much of the information provided in this article is based on presentations made at the EALTA Conference in Voss (June 2005) and at the Language Testing Research Colloquium (LTRC) in Ottawa, (July 2005)[1].

Ethics and principles of good practice

The early work of ALTE Members in this area took place in the 1990s when they addressed the question of what a code of practice might be like and what the underlying principles should be. This led to the publication of the ALTE Code of Practice in 1994, and at about the same time, the adoption of the ALTE Standards and Principles of Good Practice.

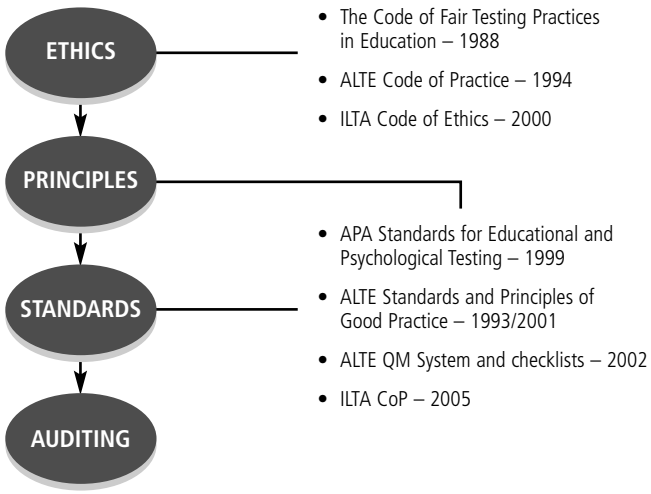
More recently a Code of Practice Working Group (CoPWG) was set up (Lisbon, 2000) to take this early work forward in light of changes to the ALTE membership and advances in the field of language testing in the 1990s. Since then it has met regularly and the main outcomes include the production of detailed documentation up-dating the principles of good practice, proposals for minimum professional standards and the implementation of appropriate checklists and procedures for monitoring those standards. The most recent work in 2004–5 has focused on the development of an ALTE auditing system.

The methodological framework which this work represents can be shown as in Figure 1. Some of the relevant reference documents are listed on the right of this figure. The Quality Profile is created in each case, by explaining how the examination meets the following minimum standards, and by providing adequate evidence.

Most of the reference literature shares an overriding aim which could be summarised as follows:

> *…. to strive for overall fairness in the testing process, always acting in good faith and avoiding negative impact, in order to provide a high quality of service to the clients and other test users.*

Figure 1: ALTE's methodological framework for auditing Good Practice



(In this case I am using the word client or customer as in various other social contexts, like the health service in the UK, for example).

In many ways the ALTE Code of Practice bears similarities to the ILTA Code of Ethics which was extensively discussed by the members of ILTA in the late 1990s and which, under the guidance of Professor Alan Davies, was published in *Language Testing Update* in 2000. The ALTE CoP and the ILTA Code of Ethics are similar in that both approaches show a concern for the broad principles of professional behaviour by those involved in language assessment and both address the kinds of actions that should be taken in order to achieve good practice, social justice and fairness for all.

The ALTE Code of Practice, however, is based around 18 broad statements covering the development of exams, the issue of results, fairness and the relationship with test takers. The focus is on the roles of the various stakeholder groups in striving for fairness – mainly focusing on the role of the test developers, but also on the roles of other stakeholders too, such as sponsors, employers, educators and the test taker themselves. This approach is appropriate for ALTE as an association of institutional Members, each of which must work with a "language testing constituency" in its own context.

The Code of Ethics, on the other hand, presents "the morals and ideals" of language testing as a profession, as set out in nine principles with annotations in order to guide good professional conduct. These points are mainly addressed to individuals who consider themselves to be professional language testers, whether they work for institutions, such as those in ALTE, or as academics

1. The poster from LTRC is available on the ALTE web-site: http://www.alte.org/quality_assurance/code/ltrc_poster.pdf

and other professionals in the field of assessment. In summary the Code of Ethics states that professional language testers should:

- Show respect for humanity and dignity
- Use judgement in sharing confidential information
- Adhere to ethical principles in research
- Avoid the misuse of their professional knowledge and skills
- Continue to develop their knowledge and to share this with others
- Share responsibility for upholding the integrity of the profession
- Strive to improve the quality of language testing and awareness of issues related to language learning
- Be mindful of obligations to society
- Consider the effects of their work on other stakeholders.

It would be difficult to argue with the sentiments expressed in either document in terms of their aspirations and emphasis on the desirability of high levels of professional conduct. Neither document, however, is designed to assist language testing practitioners in carrying out their day-to-day work of writing and administering tests, or in agreeing on what might be acceptable in terms of minimum standards for their work.

It was in order to address these last points that the ALTE CoP Working Group was set up. As a starting point for their work the group reviewed a range of available documentation, including the *ALTE Standards and Principles of Good Practice* which had been drafted in 1993. Following discussions, a new document was produced based on this which established clear principles on two dimensions:

- the process of exam development and administration
- the context and purpose of the exam being monitored (i.e. the use made of scores and results).

This was designed to provide an explicit basis for confirming the parameters of good practice and a starting point for agreeing on the standards that should be set.

The group also thought that it was useful to ensure that all ALTE Members had access to the results of other ALTE projects that were completed in the 1990s and which were intended to provide a "toolkit" for raising the quality of Members' exams. These documents are as follows (most of which were produced in a wide range of languages):

- *Checklists for Test Evaluation*, ALTE – 1993 (EU Lingua funded project)
- *Guidelines for Test Item Writers*, ALTE – 1993 (EU Lingua funded project)
- *User Guide for Examiners*, Council of Europe – 1996.
- *Glossary of Language Testing Terms*, ALTE – 1997 (EU Lingua funded project now published by Cambridge ESOL/CUP as volume 6 in the Studies in Language Testing series)
- *Code of Practice Checklists*, ALTE – Lisbon 2000.

For more details of these projects and where to obtain copies (in English) see the ALTE website: www.alte.org

The *User Guide for Examiners* has been reprinted by the Council of Europe with a new title: *Language examining and test development*. Drawing on the more extensive work contained in the *Guidelines for Test Item Writers*, it provides a practical model for examination development, including discussion relating to the design, development, monitoring, evaluation and revision of examinations.

Other works in the field which appeared in the 1990s were also referred to, and the approach which was adopted took into account the works of Messick, Bachman, Bachman and Palmer, Kane, Mislevy, and discussions of fairness derived from work by Kunnan. In establishing the essential qualities of a test for which standards should be set, reference was made to the dimension of "usefulness" or "utility". This concept is related to "overall test validity" and takes into account the context of test development and its socio-political impact in terms of the use made of scores. This can be characterised by the following features (cf. Bachman and Palmer 1996):

UTILITY/USEFULNESS = **VALIDITY** + **RELIABILITY** + **IMPACT** + **PRACTICALITY**

A Quality Management System (QMS)

Having established the principles, and having provided some practical tools to help ALTE Members improve their examination systems, the CoPWG addressed the issue of how to put the principles into practice, how improvements could be monitored and whether adequate standards were in fact being met. While most people in ALTE agreed with the principles, it was more difficult to get consensus on how the standards could be set in an equitable way, allowing for the diversity of organisations and testing practices across ALTE as a whole.

In order to address this problem and to seek consensus, it was decided that the appropriate paradigm for this activity would be that of Quality Management Systems (such as that represented by the ISO 9000 series). Quality management systems seek to improve the products and/or services of an organisation in order to meet the requirements of its customers in the most effective way, and they go about doing so in a well-planned and focused manner. There are many examples of QMS being used in Europe, including many in educational contexts, and several of these were thoroughly reviewed by the CoPWG. This was summarised and then extensively discussed by the full membership (for more details see van Avermaet et al. 2004).

Interestingly, effective Quality Management Systems usually involve a public statement as a starting point, often in the form of a Code of Practice or Charter, and also a commitment to the *change process* typically involving the following steps:

- Define your mission, role of institution, future ambitions/aims
- Assess what you currently do
- Identify areas in need of improvement
- Decide on measures of improvement and an action plan
- Carry out action to bring about improvement
- Review progress and revise plan.

In the case of ALTE, the Code of Practice already set out the public position in terms of the aspirations and ambitions of the association (as explained above), but in adopting the QMS approach, Members undertook to understand the nature of their organisations better and in so doing to involve their stakeholders in striving for improvements in quality. In effect, this involved an ongoing commitment to "change management".

In a QM system of this kind, standards are not imposed from "outside", but are established through the system itself and the procedures to monitor standards are based on awareness raising and self-assessment in the first instance. External (peer) monitoring is introduced at a later stage to confirm that the minimum standards are being met. The CoP Working Group recommended that within each ALTE organisation the following approach should be adopted:

- Establish *desired outcomes* and impacts within the organisation (aim at *continuous improvement*)
- Discuss and agree on *minimum standards*, but establish "best practice" models as long-term target
- Monitor quality through *self-assessment*
- Seek confirmation that standards are being met through *peer review* and *auditing* systems (in this case within the ALTE membership).

It is axiomatic in this approach that improvement is always possible, even where good practice may already exist. It was agreed, therefore, that the aim for all Members should be to continue to share expertise and gradually to raise standards over time, or in other words, to aim at the *best practice models* through an on-going process of development.

In order to provide a tool to raise awareness of those areas where change was necessary or perhaps urgently required, the original Code of Practice was reworked to function as Quality Management Checklists; this re-designed format reflected the four aspects of the *testing cycle* with which all ALTE Members and other test developers are very familiar:

- examination development
- administration of the examinations
- processing of the examinations (including the marking, grading and issue of results)
- analysis and post-examination review.

The revised format provided four Checklists which were put into Excel spreadsheets for ease of use and for Members to use as evaluation tools. (These checklists are available from the ALTE Website).

It has been agreed that the QMS approach should be a supportive tool and allow Members:

- to enhance the quality of their examinations in the perspective of fairness for the candidates
- to engage in negotiations with their senior management and sponsors in a process of organisational change, (e.g. to ensure that resources are made available to support on-going improvements)
- to move from self-evaluation to the possibility of external verification in order to set agreed and acceptable standards.

By proceeding in this way, and in discussing the outcomes with colleagues, ALTE Members have been made constantly aware of the different contexts in which they all work and of the various backgrounds from which the different Members come. Much discussion has taken place around the question of how to achieve a reconciliation between diversity and professional standards which are acceptable to all, and this is now continuing in the implementation of an auditing system, outlined in the following section.

Monitoring standards – auditing the quality profile

The process of setting and monitoring standards began through self-evaluation and monitoring within each ALTE member organisation. However, it was always envisaged that this would need to be supplemented by some kind of external monitoring or "auditing" system, probably based on "peer-review", whereby the ALTE Members would monitor each other. A system along these lines, designed for monitoring the standards within ALTE, is now being developed and piloted (2005) in order to carry out the task of assessing and advising Members on achieving acceptable standards.

Taking the Code of Practice and QMS Checklists into account, 17 minimum standards have been agreed to establish a Quality Profile for an exam or suite of exams (reproduced in Table 1).

The formal external scrutiny of ALTE Members' standards will therefore be the culmination of the process of establishing audited "quality profiles" across the ALTE framework of examinations. The aim now is to allow ALTE Members to make a formal, ratified claim that a particular test or suite of tests has a quality profile appropriate to the context and use of the test.

The following points need to be borne in mind:

- Different tests are used in different contexts, by different groups of test users. There is no intention to impose a single set of uniform quality standards across all ALTE Members' exams.
- Members requesting an audit of their quality systems and procedures are invited to *build an argument* that the quality standards within a test or suite of tests are sufficient and appropriate for that test or suite of tests.
- It is the *argument* which is the subject of the audit, rather than the organisation itself (which is often dealt with by other systems of regulation, e.g. ISO 9001, government regulators etc.).
- Each audit considers one test, suite of tests or testing system.
- The audit has both a consultancy and quality control role.
- The audit aims to establish that minimum quality standards are being met in a way that is appropriate to the context of a test, and also to offer recommendations towards best practice where, though quality standards are appropriate, there is still room for improvement.
- If quality standards are not being met, ALTE Members will collaborate with the audited organisation to implement an action plan aimed at working towards and ultimately reaching the quality standards.

Table 1: Minimum standards for establishing Quality Profiles in ALTE examinations (draft, summer 2005)

• TEST CONSTRUCTION	
1	The examination is based on a theoretical construct, e.g. on a model of communicative competence.
2	You can describe the purpose and context of use of the examination, and the population for which the examination is appropriate.
3	You provide criteria for selection and training of test constructors and expert judgement is involved both in test construction, and in the review and revision of the examinations.
4	Parallel examinations are comparable across different administrations in terms of content, stability, consistency and grade boundaries.
5	If you make a claim that the examination is linked to an external reference system (e.g. Common European Framework), then you can provide evidence of alignment to this system.
• ADMINISTRATION & LOGISTICS	
6	All centres are selected to administer your examination according to clear, transparent, established procedures, and have access to regulations about how to do so.
7	Examination papers are delivered in excellent condition and by secure means of transport to the authorised examination centres, your examination administration system provides for secure and traceable handling of all examination documents, and confidentiality of all system procedures can be guaranteed.
8	The examination administration system has appropriate support systems (e.g. phone hotline, web services etc).
9	You adequately protect the security and confidentiality of results and certificates, and data relating to them, in line with current data protection legislation, and candidates are informed of their rights to access this data.
10	The examination system provides support for candidates with special needs.
• MARKING & GRADING	
11	Marking is sufficiently accurate and reliable for purpose and type of examination.
12	You can document and explain how marking is carried out and reliability estimated, and how data regarding achievement of raters of writing and speaking performances is collected and analysed.
• TEST ANALYSIS	
13	You collect and analyse data on an adequate and representative sample of candidates and can be confident that their achievement is a result of the skills measured in the examination and not influenced by factors like L1, country of origin, gender, age and ethnic origin.
14	Item-level data (e.g. for computing the difficulty, discrimination, reliability and standard errors of measurement of the examination) is collected from an adequate sample of candidates and analysed.
• COMMUNICATION WITH STAKEHOLDERS	
15	The examination administration system communicates the results of the examinations to candidates and to examination centres (e.g. schools) promptly and clearly.
16	You provide information to stakeholders on the appropriate context, purpose and use of the examination, on its content, and on the overall reliability of the results of the examination.
17	You provide suitable information to stakeholders to help them interpret results and use them appropriately.

In general terms the ALTE *Procedures for Auditing* draw on approaches to auditing adopted by EAQUALS and ISO 9001 and aim to be: professional, confidential, comprehensive, impartial, consistent and supportive.

It is planned that the auditors will be appointed by ALTE but the membership as a whole will be the arbiter of decisions arising from the auditing process (e.g. through a standing committee on Code of Practice and Quality Management issues). In 2005/6, the piloting of the system will continue and further refinement will be made; future developments, including amendments which arise from the piloting, will be reported on the ALTE website.

References and further reading

AERA/APA/NCME (1985/1999) *Standards for educational and psychological testing*, Washington: AERA.

ALTE (1993) *Principles of Good Practice for ALTE Examinations* (draft manuscript).

—(1994) *Code of Practice*.

—(1998) *Handbook of Language Examinations and Examination Systems*.

ALTE/Council of Europe (1997) *Users Guide for Examiners*, Strasbourg.

van Avermaet, P, Kuijper, H, and Saville, N (2004) A Code of Practice and Quality Management System for International Examinations, *Language Assessment Quarterly* 1: 2&3. Special Issue: The Ethics of Language Assessment. Guest Editor, Alan Davies, 137–150.

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

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

ILTA (1995) *Task Force on Testing Standards*.

—(2000) Code of Ethics, *Language Testing Update*, 27, 14–22.

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

Kunnan, A J (2000) *Fairness and validation in language assessment*, Cambridge: UCLES/Cambridge University Press.

—(2004) Test Fairness, 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.

Messick, S A (1980) Test validity and the ethics of assessment, *American Psychologist*, 35(11), 1012–1027.

—(1989) Validity, in Linn, R L (Ed.) *Educational Measurement*, New York: ACE / Macmillan, 13–103.

—(1994) The interplay of evidence and consequences in the validation of performance assessments, *Educational Researcher*, 32(2), 13–23.

Mislevy, R J (1994) Evidence and inference in educational assessment, *Psychometrika*, 59, 439–483.

Mislevy, R J, Steinberg, L S, and Almond, R G (2003) On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1(1), 3–62.

Saville, N, van Avermaet, P, and Kuijper, H (2005) *Setting and Monitoring Professional Standards: a QMS approach*, paper presented at the 2nd EALTA Conference, Voss, June 2005.

Ethical issues in the testing of young learners

JULIET WILSON, EXAMINATIONS AND ASSESSMENT GROUP

Introduction

With the increase in provision of English language teaching for children has come a growing demand for assessment both in the state and private education sectors. In turn this has led to a debate within the testing and teaching community about how best to assess children's second language skills. Teachers, school owners and parents are looking for fair and accurate ways to measure the progress of their pupils. However there are also those who have fundamental reservations and concerns not only about the number of tests that children now face but also about the very existence of formal English language tests for 7–12 year olds. These may be seen as essentially undesirable and even detrimental to the learning process. This article describes Cambridge ESOL's approach to the testing of children and the ways in which we have addressed the ethical dimensions of testing children's English.

The development of the Cambridge Young Learners Tests

Cambridge ESOL responded to requests for English language assessment for children by developing the Cambridge YLE Tests. This development began in 1993 and the tests were introduced in 1997. The test development team worked closely with staff from Homerton (Cambridge University's teacher training college) to produce tests that took account of current approaches to curriculum design and pedagogy for young learners as well as children's cognitive and first language development. As the tests have now been operational for nine years and as part of Cambridge ESOL's ongoing commitment to improving the quality of its tests, we are currently undertaking a review of the Cambridge YLE Tests (further details of this can be found in *Research Notes* issue 15).

In developing the Cambridge YLE Tests, the question facing Cambridge ESOL was whether it is possible to create international English language tests for children that provide an accurate assessment but will also have a positive impact on their learning. Cambridge ESOL has a long history of producing high quality ESOL examinations for adults and teenagers but the specific characteristics and requirements of children as language learners and as test takers needed to be considered both in terms of test design and the test taking experience.

Test design

Since the experience of taking a Cambridge YLE Test may be a child's introduction to formal assessment, the impact of the tests is a key issue. Not only should the tests give a positive first impression of international testing but they should also have a positive impact on individuals, classrooms and society in general. A number of factors were therefore considered in designing the tests, described below.

The Cambridge YLE Tests test the four macro skills, with the emphasis on oral and aural skills as these are the skills that children develop first. Topic areas are chosen which are relevant to children's lives, e.g. school, food, sports and animals. Additionally, all language which is used in the tests is placed in a clear context, as this is how children process language; there are no discrete questions testing grammar. Syllabuses and wordlists are published so teachers can fully prepare children for the tests.

In the Cambridge YLE Tests, children demonstrate their understanding and learning through 'doing', for example by colouring, drawing lines or pointing. The tests are short, but with plenty of time allowed for each task as it is important that lack of time is not a source of stress during the tests. Within the test, tasks are short and varied to keep children's attention focused as children perform best when they are engaged and motivated. All tasks are based on colourful graphics. Children are less likely to feel anxious and will be able to perform to the best of their ability if the materials are attractive and fun and do not have the appearance of traditional test materials.

Overall, YLE tasks are designed to be non-threatening, fun and reflect activities that the children would do in the classroom. There is no pass or fail in the Cambridge YLE Tests and all children who take all parts of the test receive an award. The ethos of the tests is to reward the children for what they do know rather than penalise them for what they don't.

The Speaking test

Designing the YLE Speaking test presented particular challenges. This is a face-to-face Speaking test where one oral examiner assesses one child (a 1:1 format). This could be seen as potentially a stressful experience for the child. A range of formats for the test were considered. Most of the Cambridge ESOL Speaking tests are 2:2 format (i.e. two candidates and two examiners). However, young children may well have not developed the turn-taking strategies which make the paired format successful for the adult exams. Having two examiners was also considered. Overall it was felt that having two adult strangers in the room could be unnecessarily intimidating. In the 1:1 format the examiner does in fact act as the child's 'partner', demonstrating and carrying out the various Speaking test activities with the child.

Various measures are in place to ensure that the Speaking test is a comfortable experience for the child and offers conditions where they can perform to the best of their ability. Firstly, there is always an usher on hand who speaks the candidate's first language and

who ideally is someone known to him or her. The usher's duties are specifically laid out in the administration guide for Centres. The duties include ensuring that each child knows what to expect and is not over anxious about taking the test; telling each child the name of the examiner and assuring them that s/he is friendly; and accompanying them into the test room at the appropriate time and introducing them to the examiner in English before leaving.

The examiner's role is to assess the candidate's performance accurately and to ensure all candidates are treated fairly. In the *Instructions to Oral Examiners*, examiners are specifically told to 'take special care to be encouraging to the candidates.' Unlike other Cambridge ESOL examinations where examiners are asked to avoid responses such as 'good' or 'that's right', examiners for the Cambridge YLE Tests are asked to include these positive interjections. In addition, examiners are given scope within their 'script' to repeat questions that the candidate might not understand and ask back-up questions.

The design of the Speaking test means that in the initial tasks, the candidate is not required to speak at length but to respond either by pointing or placing cards on a picture. This gives the child the chance to get used to the examiner's voice before producing language themselves.

The test taking experience

Candidates are encouraged to take the Cambridge YLE Tests in their own classrooms. This means that their surroundings will be familiar and they are therefore less likely to be anxious. For the Speaking tests the examiners are asked to arrange the furniture so that the candidate sits either next to the examiner or at right angles depending on what is appropriate in the local context.

In addition to the test design and experiential factors, all of our candidates, including children taking the YLE Tests, are subject to privacy and protection arrangements.

Privacy and protection

The *Administration Instructions* for Cambridge ESOL Centres outline the policy in terms of individual privacy and protection of candidates which Centres should follow. The implications for testing children are highlighted here:

> *We would …ask our centres particularly to ensure that the appropriate steps are taken to ensure the safety of children and young learners in all matters relating to the administration of Cambridge ESOL exams.*

One of the issues which Cambridge ESOL takes very seriously is that of the suitability of those who are going to examine children. Each particular country will have its own regulations and local law. In the UK, any potential YLE examiner has to undergo a police check. Outside of the UK, it is the responsibility of the Local Secretary to ensure that anyone recruited as a YLE examiner is a suitable person. The Minimum Professional Requirements for oral examiners state that applicants should have recent experience of dealing with children either socially or professionally and must be prepared to sign a declaration that they are suitably responsible to examine children. There are also practical guidelines – YLE Speaking tests, wherever possible, should be held in rooms with an interior glass window or door, for example.

Conclusion

Cambridge ESOL takes very seriously its commitment to providing ethical and fair tests for all of our candidates and our involvement in the area of testing children has highlighted the importance of issues of privacy and protection. We continue to review the policies in place and take account of new laws and regulations, including the professional standards described in Nick Saville's article.

For the YLE Tests, as for all of our language tests, Cambridge ESOL engages with a wide range of external stakeholders to ensure that these tests are meeting the needs of test takers, administrators, parents and other interested groups. One way in which this is done is through stakeholder questionnaires or research studies; both approaches continue to be used to good effect with the YLE Tests.

Besides our own work on Cambridge YLE Tests, people external to Cambridge ESOL also consider the nature and impact of these tests. One example of this is Alison Bailey's recent review of the YLE Tests which appeared in the journal *Language Testing* and was reviewed in *Research Notes* 21. Bailey describes key features of the YLE Tests and evaluates their essential test qualities of validity, reliability, fairness (developmental appropriateness and cultural sensitivity), practicality (administration and scoring), and impact. She concluded that the Cambridge YLE tests are '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'.

For further information about Cambridge YLE Tests visit the exam information page and teaching resources accessible from www.CambridgeESOL.org

Reference

Bailey, A (2005) Test review: Cambridge Young Learners English (YLE) Tests, *Language Testing* 22 (2), 242–252.

An overview of computer-based testing

PAUL SEDDON, PROJECTS OFFICE, ESOL OPERATIONS

Introduction

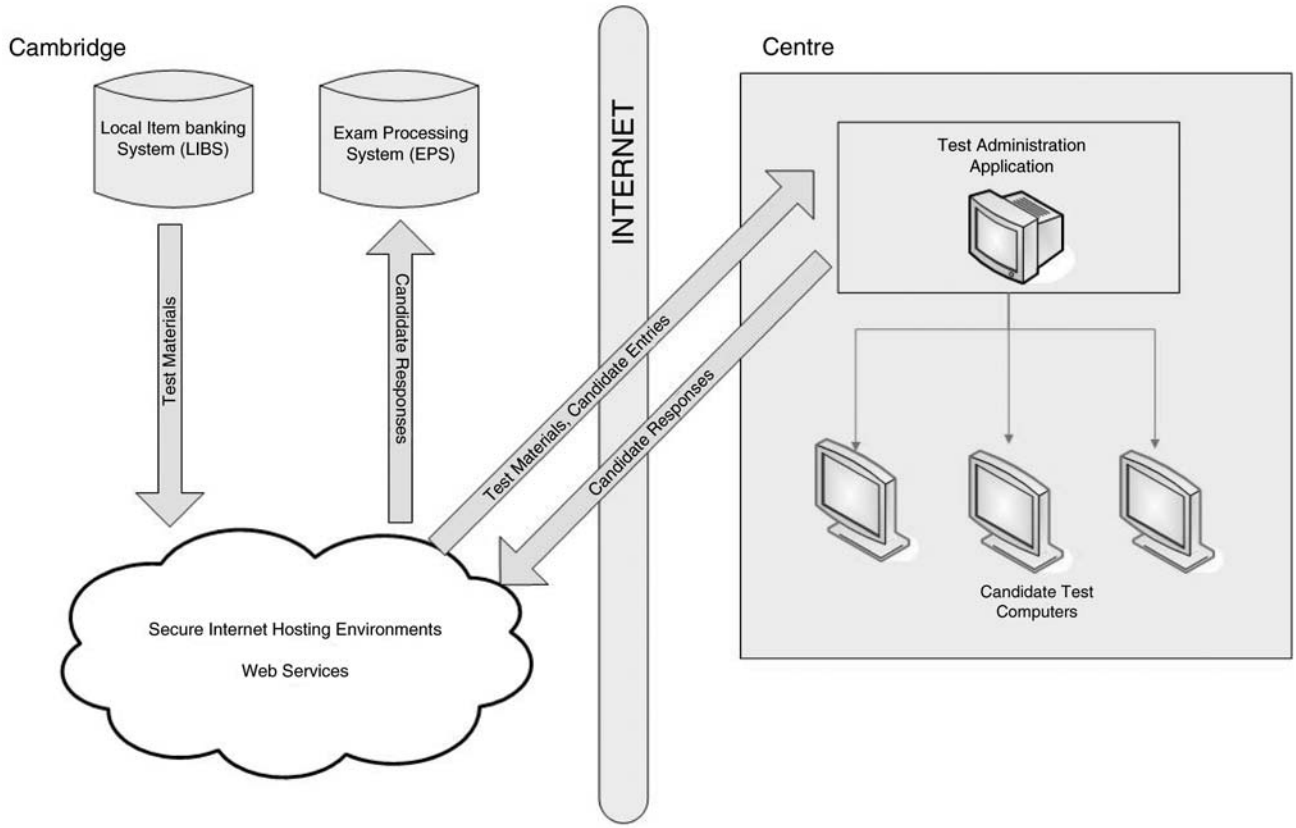
Over the last two years Cambridge ESOL, in conjunction with in-house IT development teams, has been involved in the creation of a new online test delivery engine. The engine is made up of a series of separate components each responsible for different functions collectively referred to as the Connect Framework. The framework caters not only for the delivery of computer-based tests at test venues, but also ties in to the backend systems that drive Cambridge ESOL examinations and assessments, including our Local Item Banking System (LIBS) and our Examinations Processing System (EPS), as well as Online Entries, Online Marks Capture and Online Return of Results initiatives.

Figure 1 shows Cambridge Assessment's Connect Framework. The Framework itself is generic enabling Cambridge ESOL to deliver a number of different assessments onto the same computer-based testing delivery system; the first assessment utilising this platform is a computer version of the Preliminary English Test (PET) and this will shortly be followed by other ESOL examinations.

How does it all work?

Within LIBS tasks selected for inclusion into a test version are marked up in an internationally compliant version of XML (extensible mark-up language) using specially designed mark-up tools; this enables all the tasks to be stored as an electronic test. Added to this test 'bundle' are all associated media files, audio and graphic, which in addition to the test content are then heavily encrypted and sent to web servers. At approved computer-based (CB) test centres a CB Administration application (Connect Control) is installed enabling dedicated test supervisors to login, view which test sessions are available, when they are available and which candidates have entered for the assessments. There can be as many as three sessions available for a centre on any one test date. At the specified time for the test to run, the test materials are unencrypted and test supervisors print out attendance registers and candidate login details which are distributed to each candidate. Candidates' machines are 'locked down' ensuring that no other applications can be used at the same time as the test, for example spelling and

Figure 1: Cambridge ESOL's Connect computer-based testing framework



grammar checks. Candidates log into their Connect Stations using the login details, check their sound levels, confirm their details are correct, watch a short tutorial, and then start the test when instructed.

Reading, writing and listening components are all taken on computer; the speaking component remains a face-to-face interview with examiners. During the assessments all candidate responses are automatically saved to the Connect Control computer – if a candidate's computer fails for any reason they can restart the test from where they left off. The system also caters for test interruptions, for example if a candidate needs to leave the exam room or a fire alarm requires an evacuation, then the candidate's test can be interrupted and resumed at a later time. At the end of the assessment the test supervisor simply presses a button to upload all candidate responses back to Cambridge where they are processed for marking within our Exams Processing System either automatically (for multiple-choice questions), clerically or examiner marked. From this point onwards the grading and certification processes are similar to traditional paper-based assessments.

Why do it?

The Connect Framework offers clear benefits both to the candidates and to the centre as it provides far more flexibility and frequency of test dates. At the moment there are six fixed date exam sessions for the paper-based version of the PET exam; but with the Connect Framework many more test sessions can be offered in any one year, with sessions in months not already covered by the paper-based examination. With the addition of Online Entries, the lead-in time for entries can be shortened to up to two weeks before the day of the examination providing both candidates and the centre greater opportunity to make entries much closer to the exam date. In addition the results can be returned to the candidate online within three weeks of taking the exam, therefore ensuring that a candidate can enter for an exam, take the exam and receive their results within five weeks.

For the centre there is no longer the need to receive exam materials and ensure they are kept secure, this is all done automatically. After the test all that the Centre needs to do is upload the responses back to Cambridge making the whole process faster, more robust and less error prone. Recent trialling of the assessment and its delivery mechanisms has produced a favourable reaction from both centres and the candidates involved.

A significant amount of work has already been done on computer-based testing within Cambridge ESOL in the past few years. Some of the key aspects have been reported on in previous issues of *Research Notes*, particularly in relation to IELTS (see Blackhurst 2005, Green and Maycock 2004, Maycock and Green 2005) and issue 12 reported on a range of technological innovations including electronic script management (Shaw 2003).

Read more about the rationale and development of CB PET in the following article.

References and further reading

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

Cambridge ESOL (2002) Exploring issues in the assessment of pen-and-paper/computer-based IELTS Writing, *Research Notes* 10, 21

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

Jones, N (2003) The Role of Technology in Language Testing, *Research Notes* 12, 3–4.

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

Shaw, S (2003) Electronic Script Management: towards on-screen assessment of scanned paper scripts, *Research Notes* 12, 4–8.

The development of a computer-based version of PET

ED HACKETT, EXAMINATIONS AND ASSESSMENT GROUP

Introduction

This article describes the processes involved in the development of a computer-based version of the Preliminary English Test (PET), with particular focus on the adaptation of paper-based materials for on-screen delivery. A number of studies relating to the comparability of computer-based and paper-based test versions and score equivalence have been reported in previous issues of *Research Notes* (Blackhurst 2005, Green and Maycock 2004, Thighe et al. 2001 and Jones 2000) but this article relates primarily to the design and trialling of test tasks and navigation toolbars for computer-based (CB) PET. CB PET was the first examination to use the Cambridge Assessment online delivery engine (Connect Framework), described by Paul Seddon in the previous article.

Project development

There were a number of reasons why PET was the first product to be chosen for online delivery. The candidature has been growing rapidly over the past 5 years (45% since 2000), and there has been demand for PET sessions outside the standard exam timetable, for example in July and September. Furthermore, PET is taken by a primarily young candidature, over 70% of candidates are aged 20 or under, and this is an age group likely to cope well with keyboard technology. It was also thought that the format of the PET examination would be relatively well-suited to on-screen display.

The first task in the project was to assess the suitability of PET task types for use in a computer test and to identify any potential problems and their likely impact on test design or candidate

performance. There were four key stages of development:

- feasibility study
- task design and trialling
- navigation design and trialling
- equivalence trialling.

As mentioned in the previous article, it was decided that the Speaking test would remain in the same format as for paper-based (PB) PET, i.e. a face-to-face paired interview. The only difference from the centre's point of view would be that speaking marks would be keyed into a web-based application locally and returned to Cambridge electronically, thus reducing the time needed for the turnaround of results. It was also decided at an early stage that CB PET would retain the same exam format for Reading, Writing and Listening. In addition to the task types being the same as in PB PET, candidate results would also report on the same scale. This would allow schools to follow the same preparation course for both forms of the examination.

Feasibility study, task design and trialling

The aim of the feasibility study was to look at the suitability of the tasks in the Reading and Writing and Listening components for on-screen adaptation and to propose designs for trialling. Cambridge ESOL has produced computer-based tests in CD-ROM format since 1999, for example CB BULATS (Business Language Testing Service) and QPT (the Quick Placement Test, which is marketed by Oxford University Press), and development work had already been done on CB IELTS (launched in May 2005), so a certain amount of knowledge and expertise had already been gained from the development and use of these products. A major issue in converting paper-based materials for on-screen delivery is the use of the space on the computer screen. One key difference between the majority of paper-based tests and on-screen display is aspect; most test papers are in portrait view, whereas computer screens are in landscape view. Furthermore, in a paper-based test, the candidate can view two pages of text at one time. In addition to these differences, part of the screen in a CB test is taken up with navigation buttons. This does not present a problem for discrete tasks (tasks with only one item) which can be displayed on-screen in their entirety, e.g. PET Reading Part 1 and PET Listening Part 1. An example of a discrete CB task is given in Figure 1.

So long as the font is clear and of appropriate size, candidates should not encounter problems reading the text, question and options. These task types have been successfully used in CB BULATS and other Cambridge ESOL CB products.

A key step in the feasibility study was to separate out those tasks successfully used in previous Cambridge ESOL CB products and to carry out a risk assessment on the other tasks to see if there were any particular features connected with the processing of the items in these tasks that might present a problem for on-screen display and impact on candidate performance. In addition to this, the layout of previously used task-types was reviewed to see if advances in technology presented opportunities for improvement.

Multi-item tasks, particularly reading tasks with longer texts e.g. PET Reading Parts 3 and 4, present a number of problems for on-screen display as not all the text or items can be viewed at one time. Earlier Cambridge ESOL CB tests used pagination, clicking to turn to a new page, to overcome this problem, but most websites now use scrolling to move through text. Given the linear relationship between questions and text, i.e. question 1 relating to the first piece of text and question 2 to the next piece of text, it was felt that either procedure could work, so it was decided that both methods should be trialled to elicit a preference.

One reading task, Reading Part 2, was highlighted as a potential risk. This is a multiple matching exercise in which the candidate has to match five descriptions to one of eight short texts. In the PB test, both text and items are displayed on a double A4 page spread, so the candidates are able to read the questions and texts in any order and check references in one text against another

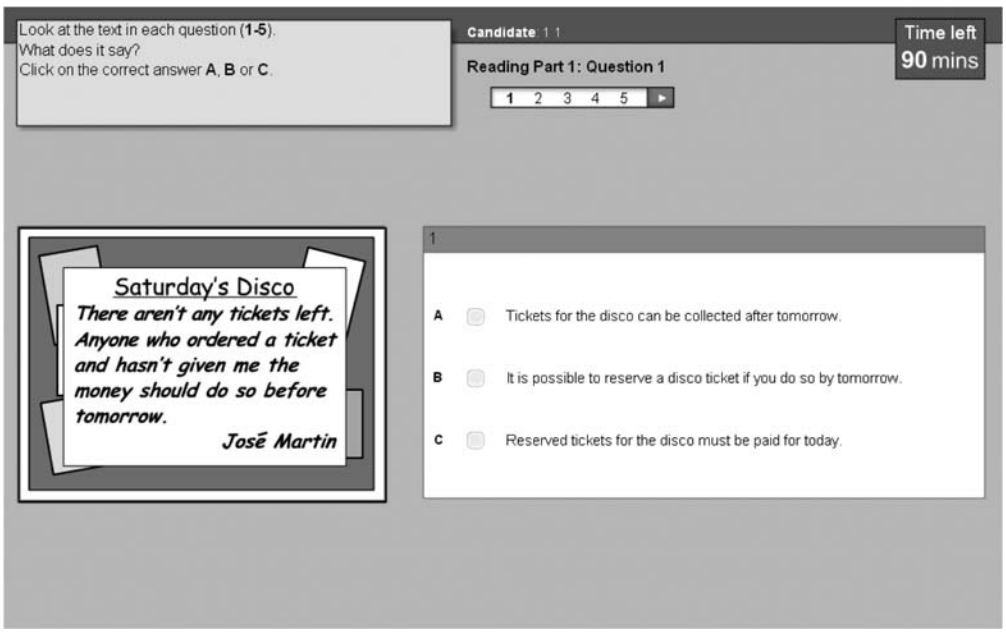


Figure 1:
CB PET Reading Part 1

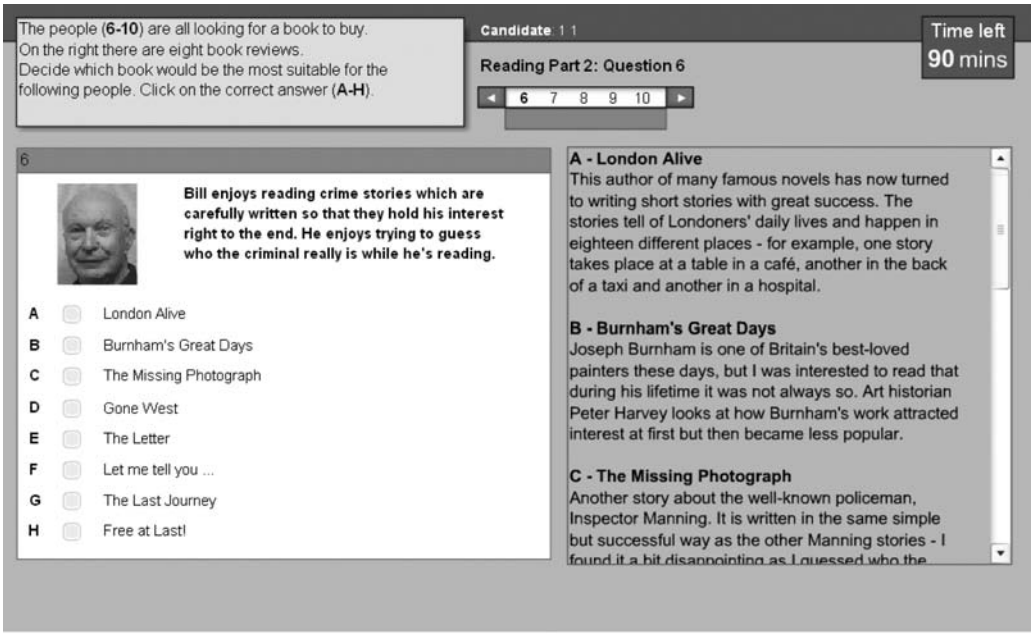


Figure 2:
CB PET Reading Part 2

simultaneously. Various options for display were looked into, including the separate scrolling of both questions and texts. However, designs for double scrolling were found to be too complex, leaving single scrolling or pagination options to be trialled. Figure 2 shows the computer-based version of Reading Part 2.

For the Writing section, the key issues were the impact of typing on candidate performance, and the effect of type-written script on examiner marking; i.e. do examiners treat typed script more harshly or leniently than handwritten script? A number of studies into this area have been carried out for CB IELTS (Thighe et al. 2001, Green and Maycock 2004), but given the different test format and candidature, it was agreed that further validation studies would be carried out for CB PET. In the Listening component, the main concern was over Part 3, a gap-filling exercise. Whilst most people can listen and takes notes simultaneously, it was felt that some candidates at this level may struggle to listen and type their answers at the same time, and that it may be necessary to allow students to make notes first, and then type in their answers.

Task trialling was carried out with a mixed nationality group of students preparing for PB PET in March 2004. Feedback from trialling was positive, with few problems being encountered by the students. Scrolling was found to be more popular than pagination as a method of moving through text, and black font on a blue background the favoured colour option for text and questions; as opposed to black on white, or white on blue. Reading Part 2 appeared to present few problems in scrolling format with one question on-screen at a time. It may be that restricting the number of questions a candidate can read at one time actually engenders a more efficient reading process and it is hoped that further research into the mechanics of reading on-screen versus on paper can be carried out in future. In writing, a preference for seeing questions one at a time was expressed for Part 1, and most candidates found typing as easy or easier than having to write by hand in Parts 2 and 3. In the listening section, candidates preferred the option of seeing all the questions at one time and a number of candidates expressed a desire to make notes first then type in Part 3.

Navigation design and trialling

Once task templates had been trialled and designs modified, the navigation system was developed more fully. The main aim in the development of the on-screen navigation toolbars was to allow candidates to progress through the test as they would in the PB format, choosing which questions to answer first and being able to return to questions at any time during the test. The design team drew heavily on standard navigation toolbars and practices used on websites. The main navigation was positioned at the bottom of the page, listing all parts of the test and this also indicated which questions had been answered. In addition to this, a part or sub-navigation was positioned above each task being attempted. This allowed easy navigation between the questions in a particular part of the test. Finally, a 'next' button appeared each time a question was answered, allowing candidates who wished to progress through the test in question order to do so. Figure 3 shows an extract from the CB Listening paper.

Navigation trialling was tested on a mixed-nationality group of Common European Framework A2/B1 level students from a UK language school in April 2004. These students were not on a PET preparation course, as it was felt that if non-PET candidates with little or no knowledge of the test format were able to successfully navigate through the test, this should not present a problem for real PET candidates. Results of navigation trialling were very encouraging, with all candidates being able to work their way through the test without instruction. A few candidates had problems with Reading Part 2, but in the focus group it emerged that they were unfamiliar with this task type. It also emerged that the rubric was not always being read, possibly due to the fact that it was of less prominence than on the PB test. The rubric box was redesigned with a clearer border and a shaded background to make it stand out more.

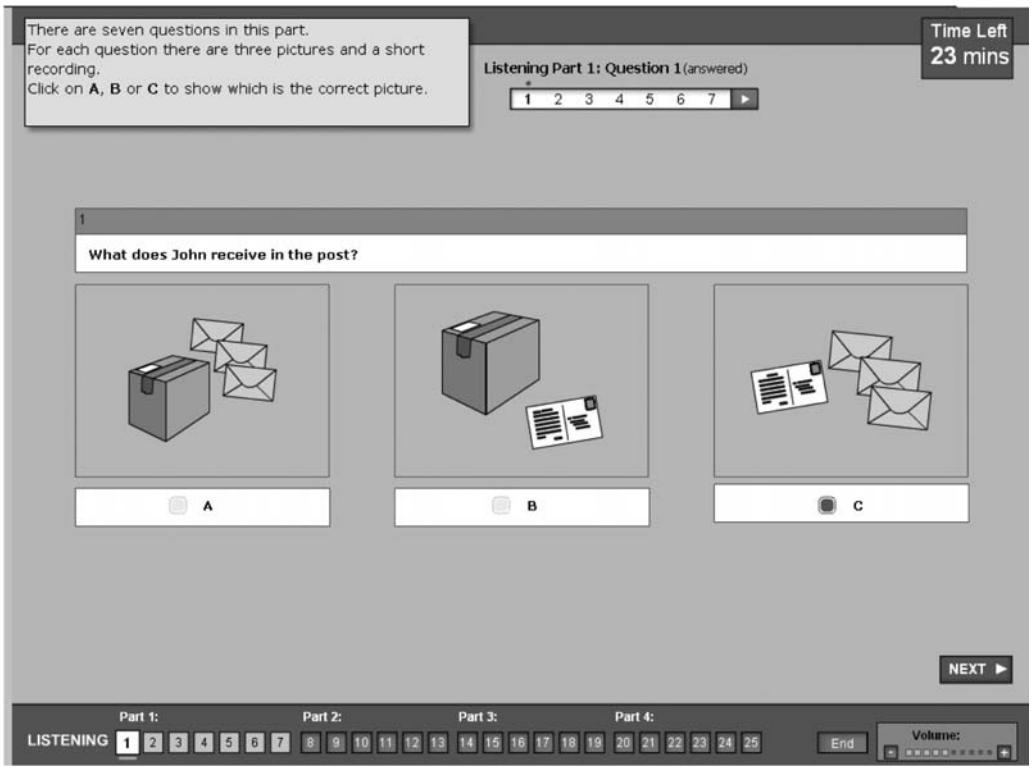


Figure 3:
CB PET Navigation Layout

Despite the relative ease with which candidates navigated through the test, it was agreed that both components: Reading and Writing, and Listening, would contain a brief tutorial, available to candidates prior to starting the main test screen. One further aspect to arise out of navigation trialling was that some candidates expressed a wish to erase earlier answers, a function which the response buttons did not allow, as once selected they could be switched to another answer, but not cleared. This functionality was added to the system ahead of equivalence trialling.

Equivalence trialling

Equivalence trialling took place in February 2005 with 190 candidates at selected centres in Ireland (with Spanish speaking students), Italy, Switzerland and the United Arab Emirates. The main aim of this phase of trialling was to establish equivalence between PB PET and CB PET, though this was also another opportunity to road-test task and navigation design and functionality. Candidates participating in trialling also took a paper-based anchor test, and were scheduled to enter for the paper-based March PET session. Analysis of results found a predicted level of score gain from the CB test to the PB test, taken 6–8 weeks later. Earlier studies in CB/PB equivalence (Jones 2000, Green and Maycock 2004, Blackhurst 2005) found no significant difference to scores gained on different formats of BULATS and IELTS tests, and it is anticipated that this pattern will be reflected for CB PET.

Equivalence trialling also gave a further opportunity to test the robustness of the online delivery engine and to get additional candidate reaction to task design and navigation usability. All candidates completed a questionnaire and a number of candidates were asked to participate in post-test focus groups. The vast majority of candidates rated the test navigation easy to use, with 96% giving ratings of 3 or above on a scale of 1 to 5, where 5 indicates total agreement. A number of specific questions relating to candidates' reactions to reading, writing and listening on computer were asked in order to gauge the general suitability of taking a test on computer as opposed to on paper. In response to the question 'Did you find reading on computer easier than reading on paper?' 46% found it easier, whereas only 25% preferred reading on paper. This perhaps reflects an increasing familiarity with on-screen reading, at home, in school or at work. Typing written answers on computer was significantly more popular than writing by hand, with 67% showing a preference for typing and only 25% expressing a preference for handwriting.

Listening was even more popular on computer, with 87% expressing a preference for listening individually on computer to listening as a group from a CD or cassette player. This is probably not surprising, as listening through headphones, with the ability to adjust the volume to your desired level, is a more personal experience. The response method chosen in Listening Part 3, where candidates have to type in a one or two-word answer, was of particular interest. Candidates were given the option of typing as they listened, or listening and making notes on paper, then typing in their answer. Two minutes were allowed at the end of the part for this. Responses were fairly evenly matched, with 42% claiming to have typed as they listened and 53% saying that they made notes then typed in their answers.

Overall, a preference for taking PET on computer was expressed by the majority of candidates. 63% preferred taking the Reading and Writing test on computer, as opposed to 20% preferring the paper-based version. For the Listening test, 83% expressed a preference for the computer version, with only 4% preferring the paper test. These results were backed up by comments made by candidates in the focus groups. Whilst there was general satisfaction with the screen layout and navigation toolbars, a few

candidates expressed a desire to be able to use a highlighting tool in the reading section, mirroring the function of underlining text on paper. The technology required to enable this function is currently being investigated for a future release of the software. Some candidates also expressed a desire to be able to control the start of each part of the Listening test. However, whilst it would be possible to cut up the sound files by part, this would remove equivalence of format with the paper-based test. In addition to measuring any impact on task difficulty, problems would arise over setting a maximum time for the length of pause between different parts and for the test itself.

Conclusion

Early indications are that CB PET appears to be well suited to a sizeable proportion of the PET candidature and it is anticipated that it will become increasingly popular with centres looking for greater flexibility and faster turnaround times. However, it is appreciated that not all centres are equipped to deliver computer-based products and some candidates will still prefer to take the paper-based version, so CB PET has been developed as an additional service rather than a replacement for traditional PB PET sessions. CB PET is being launched with a small number of European based centres in November 2005, prior to a wider word-wide rollout from March 2006.

References

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

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

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

Evaluating the impact of word processed text on writing quality and rater behaviour

STUART SHAW, RESEARCH AND VALIDATION GROUP

Background

Computers are now an established part of daily living having had a dramatic impact on both society and education (Russell and Haney 2000). The use of computers in educational institutions has become commonplace and the vast majority of students, after developing a mastery over keyboarding skills from a very early age, are now computer literate. The most common educational use of computers by students, according to Becker (1999), is for word processing. As the fundamental writing tool provided by the computer, word processors facilitate the mechanical application of placing words on to paper; revising text by substitutions, deletions, additions, and block moves; and generating presentable, aesthetically pleasing and readable final copy. MacArthur (1988) has summarised key features of word processors in the following way: word processors permit flexible editing of text, provide tidy and presentable text, and change the physical process of producing texts, replacing handwriting with typing. He has further suggested that the ease of the insertion, deletion, and overall movement of lexical units (words, sentences, paragraphs) would significantly impinge upon the processes of revision.

As the numbers of computers in educational settings have increased, various theories about how computers might benefit students' writing have proliferated, leading to investigations designed to examine whether composing on computer results in better writing. Concomitant with the increased use of computers is the impact word processed text has had and continues to have on the assessment of writing.

The word processor as a writing tool

As a composing medium the word processor has the potential for radically altering the writer's composing process and product. Several studies indicate that word processors have an effect on student writers' attitudes, the nature and characteristics of their texts, their revising activities and the stress they place on form and mechanics, and the order and the type of writing activities in which they engage (Bangert-Drowns 1993, Cochran-Smith 1991).

In the L2 context, the mechanical capabilities of word processors are especially pronounced where the physical process of planting words on paper and revisiting and revising text to a completed form, and the cognitive processes underpinning these, are more effortful and less automatised than when composing in the L1 (Jones and Tetroe 1987). Word processors may additionally help L2 writers who, perhaps more than inexperienced L1 writers, lack confidence in their ability to write in a second language (Betacourt and Phinney 1988). Phinney (1991) and Pennington (1999) contend that word processors can alleviate the anxiety certain L2 writers experience when writing the L2 script, when producing academic texts in their L2, and when writing generally. Other studies conducted with L2 writers report positive attitudes

associated with word processing (Neu and Scarcella 1991, Pennington and Brock 1992, Phinney 1991, Phinney and Mathis 1990). Akyel and Kamisli (1999), working with a group of Turkish university students writing in English, report that the use of computers improved student attitudes towards writing whilst simultaneously developing their confidence. Rusmin (1999) conducted a longitudinal study involving a small group of 27 experienced ESL writers in Hong Kong who were able to use the computer as much or as little as they wished in their written work for a course. The vast majority of participants were favourably disposed to the use of the computer and adopted it for their writing from the outset of the academic term or increasingly as the course unfolded.

Related to student attitude is self-awareness. Students who compose using computers are led to write in a self-conscious way and with greater engagement than when writing using more traditional means. Thus they tend to write more with an emancipated mind and 'less rewriting anxiety'. Greater involvement with the text might cause the writer to compose over a longer period of time thereby producing longer texts. A number of studies undertaken with L2 writers (Brock and Penington 1999, Chadwick and Bruce 1989, Pennington and Brock 1992) report that a general effect of word processing is the production of longer texts.

Rater perceptions of handwritten and word processed texts

The most immediate difference between paper and pencil and computer-based tests of writing would seem to be handwriting. Although not regarded as a specific criterion for assessment, the quality of handwriting exhibits a pronounced effect on the ease with which a rater may read and assess any text (Bull and Stevens 1979, Brown 2000, Sloan and McGinnis 1978). The introduction of computer administered direct tests of writing raises fundamental considerations regarding salience of legibility and the rating of second language writing (for a more complete treatment refer to Shaw 2003).

Intuitively, it may be expected that writing presented as hand-written text would be awarded a lower score than their type-written counterparts (Powers et al. 1994). This is a reasonable assumption given that it is rooted in a body of research which consistently reports a rater preference for tidy, neatly presented writing (Chase 1986, Markham 1976). In fact, research seems to indicate that the opposite is the case. As Bennett (2003) observes, available research tends to suggest that typed essays receive lower scores, i.e. raters are more lenient towards handwritten essays, possibly because substantive and mechanical errors stand out more, since the responses are easier to read. Arnold et al. (1990) suggested that raters give students the benefit of the doubt in situations where the handwriting is difficult to read. Moreover, rater expectations of computer derived text may be higher (Arnold, Legas, Obler, Pacheco, Russell and Umbdenstock 1990, Gentile, Riazantseva and Cline 2001).

One study that directly compared writing scores in paper and pencil tests and computer-based tests and used L2 participants found that handwritten scores were higher across L2 group classifications, while high levels of word-processing experience reduced the difference (Bridgeman and Cooper 1998). Although response format seems to have little impact on scores, Brown (2000), Shaw (2003) and Whitehead (2003) have all identified differences in the way that L2 examiners approach typed and handwritten scripts.

In their investigation of *The Praxis Series: Professional Assessments for Beginning Teachers* (an ETS test offered to test takers in both paper and computer mode), Powers, Fowles, Farnum and Ramsey (1994) found that handwritten responses received a higher score than their neatly formatted computer counterparts regardless of the mode in which the essay was originally produced. In a subsequent repetition of the experiment, Powers et al. (1994) presented raters with computer responses comprising double-spaced text (perceived length of text produced by test takers could be a contributory factor in explaining the lower scores manifest in computer generated text). Double-spacing appeared to reduce the magnitude of the effect although computer text continued to receive lower scores. Powers et al. (ibid.) provide a number of hypotheses based, in part, on research contributions from Arnold, Legas, Obler, Zpacheco, Russell and Umbdenstock (1990) which seek to explain the apparently counter-intuitive observation of lower scores for computer-generated text:

- fully edited and highly polished computer-generated products may engender higher rater expectations
- handwritten text may provoke an enhanced reader-writer relationship allowing for "*a closer identification of the writer's individual voice as a strong and important aspect of the essay*"
- poor penmanship may elicit reader sympathy prompting instances of 'benefit of doubt'
- frequently appearing longer than computer text, the construction of handwritten responses may convey to the rater a greater sense of effort on the part of the writer.

A similar study was conducted by Russell and Tao (2004) in which responses produced by students were presented for assessment in three different formats: handwritten, single-spaced 12 point word processed text, and double-spaced 14 point word processed text. Russell and Tao drew the same conclusions as Powers et al. in that handwritten responses attracted significantly higher scores. Unlike Powers et al., Russell and Tao observed that double-spaced texts (which gave the impression of being longer than single-spaced text) received lower scores than their single-spaced equivalents. Russell and Tao (2004) found little evidence that adjusting line spacing in essay texts written in computer form in an attempt to increase the perceived length consistently reduces the presentation effect. Information garnered from examiner interviews led Russell and Tao to postulate three possible explanations as to why examiners tended to award higher scores to handwritten responses. Firstly, it is easier to overlook or disregard certain mechanical textual features such as typos, uncapitalised letters and errors of punctuation when manifest as handwritten text. Secondly, raters articulate a greater sense of 'connectedness' to the writer as an individual – a possible consequence of direct exposure to the

author's personal handwriting hence the tendency to offer writers benefit of doubt. Lastly, raters are more censorious with regard to mechanical errors in typed text. Often associating computer text with a final and 'polished' version, errors of this type are construed as being symptomatic of a lack of careful proof reading – perhaps a distinguishing hallmark of novice writing. Identical handwritten errors, however, are more sympathetically interpreted and treated. The assumption on the part of the rater being that given more time the writer is able to make the necessary corrections.

The effects of the word processor on the quality of writing

Word processors may have a significant effect on writing processes with the potential to either interrupt or facilitate the cognitive processes entailed in the planning stage, the actual production of a final text, and the revision process. Herrman (1987) has argued that learning to manage the keyboard whilst writing interferes with the composing process. Conversely, Branan (1984) and Willer (1984) have each observed enhanced creative writing by learning disabled children when they received tuition in word processing.

Although research indicates that handwritten essays comprise shorter sentences (Collier and Werier 1995), contain far fewer mechanical errors (Gentile et al. 2001), appear more coherent (Russell and Haney 1997), and are generally more presentable, more formal in their tone and tend to exhibit weaker voice (Wolfe, Bolton, Feltovich and Niday 1996) than computer written essays, a number of small-scale studies have demonstrated that regular use of computers for writing over extended periods can lead to significant improvements to students' writing skills (Russell and Plati 2001). Individual studies have clearly shown that composing on a computer can increase the amount of writing students produce and the extent to which students edit their writing (Dauite 1986, Etchinson 1989, Vacc 1987), which, in turn, leads to higher quality writing (Kerchner and Kistinger 1984, Williamson and Pence 1989).

Cochran-Smith's (1991) qualitative literature review on word processing and composition in elementary classroom environments revealed that students across a large age range held positive attitudes towards word processing and appeared to be able to master a number of keyboarding strategies for use in age-appropriate writing activities. Moreover, users of word processors tend to spend more time writing and generating slightly longer, more presentable, and more technically error-free texts than when using paper. In the main, however, the review also implied that use of a word processor in and of itself does not produce better quality of writing.

Summarising prior research on word processing in language education, Wolfe and Manalo (2001) offer two somewhat intriguing possibilities. Firstly, examinees with limited word processing skills could be distracted from the writing task at hand because of the additive cognitive demands of familiarising themselves with the layout and functions of the keyboard (Dalton and Hannafin 1987, Porter 1986) and of composing on a keyboard (Cochran-Smith 1991). Secondly, despite the fact that many examinees have considerable word processing experience, surface-level changes rather than deeper, meaning-based changes might be facilitated in their writing (Hawisher 1987, Kurth 1987, Lutz 1987).

Word processing implications for planning

In the paper-based mode of composing, writers often expend considerable time and energy in intensive planning prior to writing in an attempt to obviate the need to rewrite or recopy text as a result of say, a change of mind or the correction of mistakes. Given this, it is quite conceivable that pen and paper writers may habitually write on paper without recourse to revision or with only a minimum amount of revision to avoid producing more than one draft. Conversely, the automated text-generation and revision facilities offered by word processors, in conjunction with the adaptability of text on screen, encourage a very different computer-based writing mode (Bernhardt, Edwards and Wojahn 1989, Haas 1989, Williamson and Pence 1989).

Unlike writing on paper, computer writers often begin composing immediately, soon after or sometimes before a topic is decided. Computer writers tend to plan as they write as opposed to writing to accommodate a plan (Haas 1989). This phenomenon has been documented for L2 writers as well as L1 writers (Akyel and Kasmisli 1989, Li and Cumming 2001). Planning assumes a place somewhere in the middle of the writing activity rather than at the beginning of it. The time and intensive cognitive activity that would have been involved in pre-planning is instead involved in writing itself. In other words, the clear separation of the composing stages of planning, writing and revising is radically altered in the computer context, in which planning as well as revision occurs as part of the composing process. The cognitive effort writing demands is distributed throughout the composing event in a computer approach and writing is developed more on the basis of tangible text already created than on an abstract plan. It would appear that this procedure is particularly helpful for L2 writers who have demonstrably less cognitive ability available for writing than do their L1 counterparts.

Word processing implications for revising

Research relating to revisions implies that students frequently increase the total number of changes as the available technology promotes both extensive and comparatively straightforward revisions (Cochran-Smith 1991). A number of studies have investigated differences in ways writers revise their text when using a word processor as compared to other more traditional means. Some studies have concentrated on the number of revisions made (Cross and Curey 1984, Hawisher 1986), while others have focused on the various kinds of revisions made by writers (Hawisher 1986, Kurth 1987, Willinsky 1990). Macarthur (1988) and Hult (1986) compared word processing revisions between experienced and inexperienced writers. L2 writers, for example, have been shown to revise more when composing on computer (Chadwick and Bruce 1989, Li and Cumming 2001, Phinney and Khouri 1993); to revise in a more dynamic and continuous manner (Phinney and Khouri 1993), and to spend more time revising in a

computer context, where they may 'continue revising after planned changes [have] been made' (Phinney and Khouri 1993:271).

Other research suggests that word processors tend to ameliorate the revision process (Nash 1985), that a great many revisions are almost entirely cosmetic (Womble 1984) and that fewer substantial revisions are actually made by students using word processors (Harris 1985). The participants in a study conducted by Willinsky (1989) indicated that their most frequent revising activities tended to entail minor adjustments such as spelling and grammar corrections or the addition, deletion, replacement, or re-ordering of words. Subjects using word processors reported more instances of revision than did those using typewriters or pens. Moreover, Willinsky (ibid.) reported more additions of sentences and of paragraphs – more significant changes than those observed in many of the earlier studies. Hawisher (1986) investigated the impact of word processing on the revision strategies of college students and deduced that the extent of revision did not appear to positively correlate with quality ratings and that students tend to make the same type of revisions irrespective of the writing instrument employed. Hult (1986) found that novice writers concentrate on surface revisions focusing their revision efforts at the level of word changes. However, 'expert' writers perceive the process of composition as a complete activity and as a consequence tend to make more style and content adjustments. MacArthur (1988) noted that when both experienced and inexperienced writers use word processors for composing the experienced writers tend to make more revisions.

Lutz (1987) studying a group of professional writers found that revisions were more effective when undertaken with pencil and paper or when done on hard copy. Lutz contends that deep textual revisions are discouraged when using a word processor. Computer screens expose only part of the content of a document at any one time thereby encouraging local editing but may well limit the writer's holistic perspective on, and appreciation of, the document in its entirety. Pennington (1996) argues that surface-level editing for both spelling and mechanical features is encouraged in a word processing environment, where the small size of text visible on screen may engender a particularly focused type of revision at word, phrase and sentence level. At the same time, the comparative ease with which individual words can be searched and entire portions of text excised, added, or moved implies that word processing may have a value as a macro-level revision tool. Rather than being a separate activity following the generation of a draft, revision in a computer context is closely linked to text generation.

Writers also demonstrate a tendency to make more revisions beyond the surface level. Evidence exists, for example, to suggest that word processing is more effective in stimulating meaning-level revision when aligned to a process approach to writing (Daiute 1985, Susser 1993) than when used without process support or with other computerised writing facilitation aids such as grammar checkers (Brock and Pennington 1999, Pennington and Brock 1992).

Perhaps the assumption that choice of writing instrument alone will engender different revision activities is in itself unrealistic. Hult (1986) has argued that the development of efficacious revision strategies necessitates the requirement for students being taught the importance of such functions as moving and deleting blocks of text. Kurth (1987) believes that the quality of students' writing is affected more by good instruction than by the writing tool selected.

Training raters to compensate for differential expectations

Raters can be trained to partially compensate for differential expectations they may have concerning the quality of handwritten and word processed text (Powers, Fowles, Farnum and Ramsey 1994). Powers et al. (ibid.) undertook a small-scale investigation in which participant readers were provided with modified training procedures introducing them to the phenomenon of 'presentation effect' and were additionally instructed to apply the same assessment criteria to both handwritten and computer generated responses. Whilst the amalgamation of double-spaced text and supplemental training brought about a reduction in the presentation effect, it was unable to eliminate it altogether. In a study of 60 responses transcribed into computer text and formatted in different ways, Russell and Tao (2004) provided evidence that the presentation effect can in fact be eliminated through training. Consensual agreement for scores awarded by raters to responses presented in different formats was accomplished by:

- providing descriptions of the presentation effect to raters
- discussing with raters the possible causes of the effect
- exposing raters to exemplar forms of responses that appear quite different when presented in both handwritten and computer-printed formats
- advocating that raters maintain a mental register of the frequency of mechanical errors encountered while carefully reading a response
- encouraging raters to consider thoughtfully the various factors that influence the judgements they make.

Shaw (2005) conducted a small-scale exploratory study with a group of experienced IELTS examiners at Anglia Polytechnic University (UK) in order to investigate the effect of presentation. Raters were shown a range of IELTS Academic Task 1 and Task 2 scripts in handwritten and typed format. Handwritten scripts representing varying degrees of legibility were transcribed into computer-text and formatted in five different ways: handwritten; typed double- and single-spaced; Arial 12 point and New Courier 10 point. The findings from this study can be summarised under general headings as follows, along with some of the raters' comments:

Format

- handwritten text allows raters to view a response in its entirety affording greater holistic appreciation
- rating word processed texts can be "*a liberating experience*" as raters sometimes have to work hard deciphering less legible handwritten text

- double-spacing is preferred as is Arial 12 point as it is "*clearer, bigger and easier to read*" whilst New Courier 10 point is perceived to be "*old fashioned*"
- reading typed text is considerably faster than reading poorly handwritten text: typed text looks considerably shorter than handwritten text.

Treatment of errors

- raters are comfortable treating typing and spelling errors in the same way: "*We don't have to make judgements if we treat them the same*". Moreover, they are confident in being able to distinguish between the two types of error – error repetition and keyboard positioning of letters are differentiating strategies employed by raters
- raters recognise the responsibility candidates have when it comes to ensuring that the final product is proofread: "*Many IELTS candidates are preparing for university and should be familiar with proof reading their work*"
- although handwritten responses containing interpolations (revised text or text added as an afterthought) are considered irritating they are not thought to be problematic.

Legibility

- raters give 'benefit of doubt' where legibility is an issue
- raters reported that their marking is helped by reading aloud poorly handwritten responses
- raters tend to scrutinise less legible scripts more often than legible ones – the first reading being a deciphering exercise with subsequent readings focusing on assessment criteria
- raters are less inclined to pore over typed text containing many typographical errors: "*Lots of typos have a more negative affect than lots of handwritten ones … I just cannot be bothered to read through all the typed ones*"
- raters are conscious that they might be influenced by neat and tidy handwriting. Wherever possible, therefore, they attempt to compensate: "*I must not think this is going to be poor if the handwriting is poor*".

Conclusion

Despite some of the tangible advantages of the computer over pen and paper composing in regard to flexibility, automation, and cognitive demands, the results of research on the quality of writing generated in a computer context are not all entirely favourable, as only some studies have yielded beneficial effects for student compositions produced by word processing in contrast to pen and paper (Pennington 1996). From the L2 perspective, a mixed pattern of findings have been reported. In some studies, word processing gives writers an advantage in terms of the quality of their writing (Lam and Penington 1995, McGarrell 1993), while in others, word processing appears to offer no advantage over pen and paper (Benesch 1987, Chadwick and Bruce 1989).

Fairness is an important consideration in the field of educational assessment. Therefore, it is important to ensure that scores obtained from both paper and pencil tests and computer-based tests are comparable and thus valid. Making raters aware of this tendency to downgrade word-processed essays has been shown to be, to some extent, an effective strategy. Whilst training reduces the effect it does not always eliminate it. Whether students are all required to produce a response to direct tests of writing in the same mode or raters are compelled to make fair evaluations of essays produced in different modes, ensuring equity in essay assessments will require research and further attention. The effect training has on reducing the presentation effect needs to be explored further by replicating, on a larger sample and on larger groups of raters, the work of Russell and Tao (2004). If subsequent trials offer evidence for the eradication of the presentation effect then a major barrier to providing test takers and students with the option of writing responses to composition-type questions on computer may be removed.

References and further reading

Akyel, A and Kamisli, S (1999) Word Processing in the EFL classroom: Effects on writing strategies, attitudes, and products, in Pennington, M C (Ed.) *Writing in an electronic medium: Research with language learners*, Houston: Athelstan, 27–60.

Arnold, V, Legas, J, Obler, S, Pacheco, M A, Russell, C and Umbdenstock, L (1990) *Direct writing assessment: A study of bias in scoring hand-written v. wordprocessed papers*, Unpublished manuscript, Rio Hondo College, Whittier, CA.

Bangert-Downs, R L (1993) The word processor as an instructional tool: A meta-analysis of word processing in writing instruction, *Review of Educational Research*, 63, 69–93.

Becker, H J (1999) *Internet use by teachers: conditions of professional use and teacher-directed student use*, Irvine, CA: Centre for Research on Information Technology and Organisations.

Benesch, S (1987) *Word processing in English as a second language: A case study of three non-native college students*, paper presented at the conference on College and Composition, Atlanta, GA (ERIC Document No. ED 281383)

Bennett, R E (2003) Inexorable and Inevitable: The Continuing Story of Technology and Assessment, *The Journal of Technology, Learning, and Assessment*, 1 (1).

Bernhardt, S A, Edwards, P G and Wojahn, P R (1989) Teaching college composition with computers: A program evaluation study, *Written Communication*, 6, 108–133.

Betacourt, F and Phinney, M (1988) Sources of writing block in bilingual writers, *Written Communication*, 5, 461–478.

Branan, K (1984) Moving the writing process along, *Learning*, 13, 221.

Bridgeman, B and Cooper, P (1998) Comparability of Scores on Word-Processed and Handwritten Essays on the Graduate Management Admissions Test, paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA, 13–17 April 1998.

Brock, M N and Pennington, M C (1999) A comparative study of text analysis and peer tutoring as input to writing on computer in an ESL context, in Pennington, M C (Ed.) *Writing in an electronic medium: research with language learners*, Houston: Athelstan, 61–94.

Brown, A (2003) Legibility and the Rating of Second Language Writing: An Investigation of the Rating of Handwritten and Word-processed IELTS Task Two Essays, in *IELTS Research Reports* Vol. 4, Canberra: IDP Education Australia.

Bull, R and Stevens, J (1979) The effects of attractiveness of writer and penmanship on essay grades, *Journal of Occupational Psychology*, 52, 53–59.

Chadwick, S and Bruce, N (1989) The revision process in academic writing: From pen and paper to word processor, *Hong Kong Papers in Linguistics and Language Teaching*, 12, 1–27.

Chase, C I (1986) Essay test scoring: Interaction of relevant variables, *Journal of Educational Measurement*, 23, 33–42.

Cochran-Smith, M (1991) Word processing and writing in elementary classrooms: A critical review of related literature, *Review of Educational Research*, 61, 107–155.

Collier, R and Weirer, C (1995) When computer writers compose by hand, *Computers and Composition*, 12, 47–59.

Cross, J A and Curey, B J (1984) *The effect of word processing on writing*, paper presented at the Mid-Year Meeting of the American Society for Information Science, Bloomington, IN (ED 247 921).

Dalton, D and Hannafin, M (1987) The effects of word processing on written composition, *Journal of Educational Research*, 50, 223–228.

Dauite, C (1985) *Writing and computers*, Reading, MA: Addison-Wesley.

—(1986) Physical and cognitive factors in revising: insights from studies with computers, *Research in the Teaching of English*, 20 (2), 141–159.

Gentile, C, Riazantseva, A and Cline, F (2001) *A comparison of handwritten and word processed TOEFL essays*, Princeton, NJ: Educational Testing Service.

Haas, C (1989) How the writing medium shapes the writing process: Effects of word processing on planning, *Research in the Teaching of English*, 23, 181–207.

Hawisher, G E (1987) The effects of word processing on the revision strategies of college freshman, *Research in the Teaching of English*, 21, 145–159.

Hermann, A (1987) *Research into writing and computers: Viewing the gestalt*, paper presented at the Annual Meeting of the Modern Language Association, San Francisco, CA (ED292094).

Hult, C (1986) *The computer and the inexperienced writer*, paper presented at the Annual Meeting of the Conference on College Composition and Communication, New Orleans, (ED 271772).

Jones, S and Tetro, J (1987) Composing in a second language, in Matsuhashi, A (Ed.) *Writing in real time: Modelling production processes*, Norwood, NJ: Ablex, 34–57.

Kerchner, L B and Kistinger, B J (1984) Language processing/word processing: Written expression, computers, and learning disabled students, *Learning Disability Quarterly*, 7 (4), 329–335.

Kurth, J R (1987) Using word processors to enhance revision strategies during student writing activities, *Educational Technology*, XXVII, 13–19.

Lam, F S and Pennington, M C (1995) The computer vs. the pen: A comparative study of word processing in a Hong Kong secondary classroom, *Computer-Assisted Language Learning*, 7, 75–92.

Li, J and Cumming, A (2001) Word processing and second language writing: A longitudinal case study, *International Journal of English Studies*, 1 (2), 127–152.

Lutz, A J (1987) A study of professional and experienced writers revising and editing at the computer and with pen and paper, *Research in the Teaching of English*, 21 (4), 398–421.

MacArthur, A C (1988) The impact of computers on the writing process, *Exceptional Children*, 54 (6), 536–542.

—(1996) Using technology to enhance the writing processes of students with learning disabilities, *Journal of Learning Disabilities*, 29(4), 344–354.

McGarrell, H M (1993) *Perceived and actual impact of computer use in second language writing classes*, paper presented at the Congress of the Association de Linguistique Appliquee (AILA), Frije University, Amsterdam, August 1993.

Markham, L R (1976) Influences of handwriting quality on teacher evaluation of written work, *American Educational Research Journal*, 13, 277–83.

Nash, J (1985) Making computers work in writing class, *Educational Technology*, 25, 19–26.

Neu, J and Scarcella, R (1991) Word processing in the ESL writing classroom: A survey of student attitudes, in Dunkel, P (Ed.) *Computer-Assisted language learning and testing: Research issues and practice,* New York: Newbury House/Harper Collins 169–187.

Pennington, M C (1999) The missing link in computer-assisted writing, in Cameron, K (Ed.) *CALL: Media, design, and applications*, Lisse: Swets and Zeitlinger, 271–292.

Pennington, M C and Brock, M N (1992) Process and product approaches to computer-assisted composition, in Pennington, M C and Stevens, V (Eds), *Computers in Applied Linguistics: An international perspective*, Clevedon, UK: Multilingual Matters, 79–109.

Phinney, M (1989) Computers, composition, and second language teaching, in Pennington, M C (Ed.) *Teaching languages with computers: The state of the art*, La Jolla, CA: Athelstan, 81–96.

—(1991) Word processing and writing apprehension in first and second language writers, *Computers and Composition*, 9, 65–82.

Phinney, M and Khouri, S (1993) Computers, revision, and ESL writers: The role of experience, *Journal of Second Language Writing*, 2, 257–277.

Phinney, M and Mathius, C (1990) ESL student responses to writing with computers, *TESOL Newsletter*, 24 (2), 30–31.

Porter, R (1986) Writing and word processing in year one, *Australian Educational Computing*, 1, 18–23.

Powers, D E, Fowles, M E, Farnum, M and Ramsey, P (1994) Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays, *Journal of Educational Measurement* 31 (3), 220–233.

Rusmin, R S (1999) Patterns of adaptation to a new writing environment: The experience of word processing by mature second language writers, in Pennington, M C (Ed.) *Writing in an electronic medium: Research with language learners*, Houston: Athelstan, 183–227.

Russell, M and Haney, W (1997) Testing writing on computers: results of a pilot study to compare student writing test performance via computer or via paper-and-pencil (ED 405 359).

—(2000) Bridging the Gap Between testing and Technology in Schools, *Education Policy Analysis Archives*, 8 (19).

Russell, M and Plati, T (2001) Effects of Computer Versus Paper Administration of a State-Mandated Writing Assessment, *Teachers College Record,* January 21, 2001.

Russell, M and Tao, W (2004) The Influence of Computer-Print on Rater Scores, *Practical Assessment, Research and Evaluation*, 9 (10), 1–14.

Shaw, S D (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.

—(2005) The impact of word processed text on rater behaviour: a review of the literature, Cambridge ESOL internal report no. 670.

Sloan, C and McGinnis, I (1978) The effect of handwriting on teachers' grading of high school essays, *Journal of the Association for the Study of Perception*, 17 (2), 15–21.

Susser, B (1993) ESL/EFL process writing with computers, *CALL Journal*, 4 (2), 16–22.

Vacc, N N (1987) Word processor versus handwriting: A comparative study of writing samples produced by mildly mentally handicapped students, *Exceptional Children*, 54 (2), 156–165.

Whitehead, R (2003) Issues in the assessment of pen-and-paper and computer-based IELTS writing tasks, final project report, Cambridge ESOL.

Willer, A (1984) Creative writing with computers: What do elementary students have to say? *Computers, Reading and Language Arts*, 2(1), 39–42.

Williamson, M M and Pence, P (1989) Word processing and student writers, in Britton, B and Glynn, S M (Eds), *Computer writing environments: Theory, research, and design* (pp.93–127). Hillsdale, NJ: Lawrence Erlbaum.

Willinsky, J (1990) When university students word process their assignments, *Computers in the Schools*, 6, 83–96.

Wolfe, W, Bolton, S, Feltovic, B and Niday, M D (1996) The influence of student experience with word processors on the quality of essays written for a direct writing assessment, *Assessing Writing*, 3 (2), 123–147.

Wolfe, E W and Manalo, J R (2001) *An investigation of the impact of composition medium on the quality of scores from the TOEFL writing section: A report from the broad-based study*, Princeton, NJ: ETS.

Womble, G G (1984) Do word processors work in the English classroom? *The Education Digest*, 50, 40–42.

Current research and development activities

Asset Languages – grading of pilot sessions

Since winning the DfES contract in October 2003 the pace of development has been hectic, and the first half of 2005 saw live pilot administrations of French, German and Spanish at Breakthrough, Preliminary and Intermediate stages. Piloting enables us to trial processing systems, and to collect reactions from centres to the tests themselves. It is also important that grading should be accurate, given that candidates receive real certificates for their efforts.

For Reading and Listening grading depends on an item-banking approach, where pretesting determines the difficulty of test tasks. Pretesting has enabled the vertical link between stages to be estimated reasonably well for most languages. Teachers' estimates of candidates' National Curriculum levels collected during pretesting also provided an indication of the target standard. Thus the grading of these skills could already proceed quite well on the basis of statistical methods. The subjectively-marked skills of Writing and Speaking present different problems, as grading these depends on standardising the judgements of raters, both across stages and across languages. Several standardisation events had addressed these issues; nonetheless, the provisional grading of Speaking raised concerns that one language was being marked more severely than the others. An expert review confirmed this, and led to an adjustment of some grades. Cross-language standardisation is a key issue in the European context, in relation to the Common European Framework of Reference. Further development of methodology for Asset will feed into the European enterprise.

Meanwhile work proceeds on the next five languages: Chinese, Japanese, Urdu, Panjabi, Italian. Tasks are currently being pretested, with the first live tests scheduled from late 2005 through 2006.

For more information on Asset visit www.assetlanguages.org.uk

CBT Research and Development

As discussed in detail by Paul Seddon and Ed Hackett in this issue, a computer-based version of the Preliminary English Test (PET) is expected to go live in several centres across Europe in November 2005. This is part of the wider development of a generic online test delivery engine (Connect Framework), which is also being used to administer Asset Languages assessments and ESOL Skills for Life, and will eventually be rolled out to a range of Cambridge ESOL examinations.

As mentioned by Andrew Blackhurst in issue 21 of *Research Notes*, the new computer-based version of IELTS has been available as a live test in a number of centres since May 2005. Regular feedback is being received after each test date and the response from both test takers and test administrators has thus far been very positive.

Recent research in the field of Computerised Adaptive Testing has addressed the issue of test re-test reliability and progress testing. Findings indicated that differences in performance between first and second attempts at CB BULATS fell broadly within the margin of error we would expect from the standard error of the difference between two BULATS tests. There was a significant increase in the average overall score between administrations taking place before and after a period of language training indicating that, in general, learning gains took place. Few candidates performed substantially worse in the second test than in the first, but the occasions where this did happen may largely be explained in terms of features of the measurement error inherent in any test. Improvement was most evident at the lowest levels, illustrating the effect of the statistical phenomenon of regression to the mean. Further articles on CB tests will appear in future issues of *Research Notes*.

ESOL staff seminar programme

Each month Cambridge ESOL staff have the opportunity to attend seminars on a range of language testing and education related topics, given by colleagues and external speakers. The eight sessions which have been run this year are reported on below.

Training tutorials

In January three internal speakers led training workshops on the skills needed by Cambridge ESOL staff in order to communicate our work in the public domain, i.e. giving presentations to internal and/or external audiences; writing contributions for internal publications (*Cambridge First, Research Notes*) and publishing articles in external venues (teacher newsletters, academic journals). These workshops outlined the breadth of venues where Cambridge ESOL staff publish and present their work and were well received by the staff who attended.

Tensions between the world of humanistic language teaching and the world of testing

In February Mario Rinvolucri (Pilgrims UK) led a workshop on humanistic language teaching and the world of language testing. The workshop opened with a paired exercise in which Person A had a conversation with Person B on a topic of B's choice: A's focus of attention was B's tempo (or speed of speech). This voice pacing activity, taken from NLP (Neuro Linguistic Processing), is a useful element in interlocutor examiner training as well as in candidate preparation for oral exams. We then worked in small groups and looked at ways in which non-neurotic test-takers lower their stress levels by what they do before a test and what they take into the exam with them. The third warmer activity focused on any doubts we had entertained, either at macro or micro level, when thinking about exams and tests.

Mario spoke to the group in Latin for a moment or two to illustrate that, for him, Latin was not a language but simply a language system, a complex grouping of forms unrelated to his emotional, relational life. Mario suggested that real language is a state of BEING and that is it always relational. Language is something well beyond HAVING certain phonological, grammatical and lexical components. Mario HAS plenty of Latin (8 years of collecting Meccano pieces) but IS not in Latin, even at a very elementary level.

We then looked at three propositions:

Proposition 1: Self-testing is central to all language production

To illustrate this claim Mario tried to uncover the exact sensory process by which a member of staff self-corrected in mother tongue when suddenly back in Bulgaria after years in the UK. It was clear that, though each person "monitors" their speech differently, we all self-check in L1, in L2, in L3 etc.

The problem, Mario suggested, arises when testing becomes a social act, beyond the boundaries of the self. This happens when parents correct, when teachers correct and when examiners correct. While self-testing happens mostly below the level of awareness, and does not arouse defences, external correction can threaten a person and may be resisted.

Proposition 2: The "Quartet" oral is a relatively complex event

This was illustrated by us listening to an experienced CPE oral examiner explaining how affective factors could sometimes make the oral test anything but a level playing field for certain candidates.

Proposition 3: Certain powerful communicative features lie outside the mark scheme of even advanced language testing

A member of staff took Mario through the first three minutes of a CPE oral where he had to speak about a picture. The group were told that one feature of his speech would be "wrong" and that they were to decide how important this feature was. Mario then spoke at a low pitch level typically associated with a certain L1 group of learners of English. There was a difference of opinion as the audience assured Mario that his wrong pitch would not count against him in the exam whilst Mario's contention was that speaking a language at the wrong Hertz levels for that language was a permanent oral feature that strongly affects communicational effectiveness.

As background reading Mario gave out an article that had appeared in the *EA Journal*: The Strange World of EFL Testing – Has psychology no place in testing (Vol. 20 no 2).

This seminar provided food-for-thought for Cambridge ESOL staff and reminded us of some of the other considerations and assumptions candidates bring to language tests.

Pretests without Protests

In March two internal speakers spoke about 'Pretests without Protests'. The commissioning of exam questions takes place at least three years before they go live. Pretesting plays a central part in the investment of objective language items into the live item bank. This seminar described some of the challenges involved in the administration of Pretests and also illustrated the criteria determining which items make it through to the completed exam.

Global English: the next stage

In April we were visited by David Graddol (director of the English Company (UK) Ltd) who spoke about 'Global English: the next stage'. David Graddol has authored several publications on the nature and implications of language change, including an influential 'think tank' report on trends in global English –

The Future of English? – produced in 1997 for the British Council. David spoke about the results of new research commissioned by the British Council which explores the likely global trends in English in the next decade. This presentation was extremely important and outlined the ways in which English is growing in usage throughout the world.

Issues in the teaching, learning and testing of children

May's seminar was given by Professor Lynne Cameron (School of Education, University of Leeds) and internal speakers on the Cambridge Young Learners English Tests (YLE). The YLE Tests are well-established in many parts of the world and are taken by around 400,000 candidates annually. Lynne Cameron described the 'complex system' which characterises the learning child in the language classroom and our own YLE team discussed some of the implications this has for our tests for young learners.

Additions to the Cambridge ESOL 'family'

June's seminar informed staff of new ESOL products and others under development in a range of new areas. The English language teaching and testing world continues to evolve. We have seen the development in recent years of the Adult ESOL Curriculum in the UK and the possibility of language tests for citizenship. There is also a growing demand for tests with a more specialised focus – to suit the needs of key professional domains in business and education. Colleagues described the implications these developments have for Cambridge ESOL and the way we are responding through our new test development programme.

Our UK Development Manager (responsible for promoting all of Cambridge ESOL's products to the UK market) began with an overview of Skills for Life. Following this, TKT was described. This is the newest teaching award which is growing in popularity and has already been taken in five countries. The third speaker described the forthcoming ILEC test and underlined some of the developmental challenges colleagues have faced to date. Finally, the work of the New Product Development Group was outlined, including the areas in which Cambridge ESOL is hoping to expand our product range in the coming years.

The E Factor – innovative options for assessment

July's seminar was given by Patrick Craven from OCR who began with an overview of the current work of OCR in relation to e-assessment. This focused on the drivers and constraints affecting assessment design, dispelling some common myths surrounding e-assessment and detailing the benefits of e-assessment. He then went on to outline the two core strands in OCR's assessment programme: e-testing and e-portfolios. The second half of the seminar provided an opportunity to compare activities within our two business streams and to demonstrate how sharing knowledge and resources would be mutually beneficial.

The role of language corpora in ELT publishing

In September three speakers from Cambridge University Press described and demonstrated their use of the Cambridge Learner Corpus (CLC) and Cambridge International Corpus (CIC) in the development of CUP's English language teaching publications.

Ann Fiddes introduced both corpora and gave examples of how these collections of written and spoken native speaker and learner English are used by a major publisher. Patrick Gillard then demonstrated searching the CIC and discussed its use in Cambridge ELT dictionaries. Finally, Annette Capel gave an author's view of how she has used corpora to help write course books for KET, CPE and IELTS examinations. In the subsequent discussion Cambridge ESOL staff raised issues such as the relationship between the frequency of occurrence of grammatical patterns in a corpus and their proficiency level and teachers' awareness of different levels of proficiency.

We look forward to more informative and thought provoking seminars throughout the remainder of this year, which will be reported in future issues of *Research Notes*.

Recent publications of interest

As discussed elsewhere in this issue, high stakes test providers such as Cambridge ESOL need to be concerned with the ethical dimension of testing in terms of the impact of a test on individuals and society; there must also be an emphasis on social values and social consequences in any consideration of the validity of test scores. Over the past 10–15 years Cambridge ESOL has been proactive in investigating various dimensions of the impact of our own tests and we have supported the work of others in this field by sharing data and/or instrumentation or by helping to publish research findings. The latest volume to appear in the *Studies in Language Testing* series is the first of several volumes to focus on washback and impact studies. Volume 21 – *Changing language teaching through language testing: a washback study* – presents a study into the impact on English teaching and learning in Hong Kong secondary schools following introduction in 1996 of a high stakes public examination. The study was conducted by Liying Cheng for her PhD. An edited version of the series editors' note for Volume 21 is given below. A second volume – *The impact of high-*

stakes testing on classroom teaching: a case study using insights from testing and innovation theory – is currently in press; and a third volume – *Impact theory and practice: studies of the IELTS test and Progetto Lingue 2000* – is due to appear early in 2006. Publication of these volumes over the coming months should enrich our understanding of this relatively under-researched area of test validity and should make more accessible the research methodologies for investigating it.

Studies in Language Testing – Volume 21

Test impact is concerned with the influence of a test on general educational processes and on the individuals who are affected by the test results. It is recognised that examination boards have a major impact on educational processes and on society in general because their examinations often have widespread recognition and 'cash in' value. Washback is an important element of test impact. While impact may occur at a 'macro' or social and institutional level, washback occurs at the 'micro' level of the individual participant (primarily teachers and students).

There is now a clear consensus on the need for a concern with, if not agreement on, the effects of what has been termed 'washback/backwash'. Washback is considered a 'neutral' term which may refer to both (intended) positive or beneficial effects and to (unintended) harmful or negative effects; it is broadly defined as the effect of a test on teaching and often also on learning. It has been associated with effects on teachers, learners, parents, administrators, textbook writers, classroom practice, educational practices and beliefs and curricula although the ultimate effects on learning outcomes should perhaps be the primary concern.

Given that language teachers have to equip students with the skills that tests are intended to provide information about, it seems likely the closer the relationship between the test and the teaching that precedes it, the more the test is likely to have washback on both staff and students. Some authors caution that although the test may influence the content of teaching this may not be uniformly positive and, more critically, tests may have little impact on methodology, how teachers teach. Liying Cheng found such a situation following the exam reforms in Hong Kong but her research clearly indicates that if adequate training for teaching the new test is not provided we should hardly find it surprising that old methodologies persist. The same is true in the Sri Lankan washback study described by Dianne Wall (to be published as SiLT Volume 22) where additionally a debilitating civil war was hardly conducive to change.

Volume 21 looks at the impact of the 1996 Hong Kong Certificate of Education in English (HKCEE), a high stakes public examination, on the classroom teaching of English in Hong Kong secondary schools. Liying Cheng investigates the effects from the decision-making level of the Education Department (ED), the Curriculum Development Committee (CDC), and the Hong Kong Examinations Authority (HKEA), down to the classroom levels of teaching and learning, with reference to aspects of teachers' attitudes, teaching content, and classroom interaction.

The study addresses the following research questions:

(1) What strategies did the HKEA use to implement the examination change?

(2) What was the nature and scope of the washback effect on teachers' and students' perceptions of aspects of teaching towards the new examination?

(3) What was the nature and scope of the washback effect on teachers' behaviours as a result of the new examination in relation to:
   a. Teachers' medium of instruction, teacher talk, teaching activities
   b. Materials used in teaching, aspects of lesson planning
   c. Assessment and evaluation in relation to their teaching.

Despite widespread lip service to the mantra of 'washback' in the international testing community, until recently only a limited number of research studies have been undertaken to study the effects of high stakes language tests on teaching and learning and even fewer were based on samples as adequate as the one employed in this study in Hong Kong.

An important strength of Liying Cheng's work is the use she made of both quantitative and qualitative methods to investigate these effects. A balanced combination of quantitative and qualitative research methods is employed to explore the meaning of change in the Hong Kong context as a result of the new examination. Phase I utilised interviews, observation, and initial surveys of teachers and students. Phase II involved two parallel survey studies of teachers and students. The major research methods used in Phase III were classroom observations and follow-up interviews. The triangulation of the methodology (multi-method methodology) and inclusion of comparable student and teacher data is of interest to all those contemplating research in this area.

The overt aim of the HKEA, in introducing the examination, was to bring about positive washback effects on teaching and learning in schools. However, the study shows the washback effect of the new examination on classroom methodology to be limited in many respects although the content of lessons shows marked change. Of particular interest is the identification of washback intensity (potential areas in teaching and learning that experience more washback effects than others within the given context of the study).

This volume will be of particular relevance to language test developers and researchers interested in the consequential validity of tests; it will also be of interest to teachers, curriculum designers, policy makers and others in education concerned with the interface between language testing and teaching practices/ programs.

Conference reports

AILA 2005

The World Congress of Applied Linguistics (AILA) took place in Madison, Wisconsin, between 24–29 July 2005. Cambridge ESOL staff organised two symposia and presented a further paper at this event, which was hosted by the American Association for Applied Linguistics (AAAL).

Vocabulary in teaching and testing contexts

Fiona Barker and Lynda Taylor took part in a symposium entitled 'Vocabulary in teaching and testing contexts: insights from corpus analysis'. This symposium included six presentations on corpus-informed research into lexical items, phraseology and formulaic sequences in English and suggested how such insights can inform teaching and testing practices. Responses to the papers were provided by James Purpura (Teachers College Columbia, New York) and Lynda Taylor (Cambridge ESOL).

Paul Thompson (Reading University) opened the symposium with a paper on the lexis of English for Academic Purposes in the UK context, followed by Susan Hunston (Birmingham University) who spoke about the implications of meaning and phraseology for language learning. Norbert Schmitt (Nottingham University) presented on the links between formulaic sequences and vocabulary teaching. Jim Purpura reflected on the key issues raised by the papers and the audience had the opportunity to respond to the issues raised in the first half.

The second half of the symposium moved from describing vocabulary for teaching and learning purposes to measuring aspects of learner language and the implications for assessing vocabulary. Alan Tonkyn (Reading University) described ways of measuring lexical range in spoken learner language, followed by Fiona Barker (Cambridge ESOL) who spoke about how a corpus of written learner English can reveal insights about vocabulary at different proficiency levels. The final speaker was John Read (Wellington University, New Zealand) who presented how the lexical dimension of the IELTS test can be measured. Lynda Taylor drew together the threads raised by the second set of papers and there was some audience discussion.

The symposium raised awareness of the ways in which corpora can be used to inform insights about related aspects of vocabulary and suggested ways in which these insights can inform teaching and testing practices. It is hoped that the papers within this symposium will be published in some form in the future.

The Big Tests: Intentions and Evidence

This symposium brought together three of the largest high-stakes tests in the world – College English Test (CET), IELTS and TOEFL. The purpose was to demonstrate why English language tests matter so much and to begin to establish a common set of standards or expectations for all high-stakes tests in this context. Firstly Jin Yan, Nick Charge and Mary Enright presented on CET, IELTS and TOEFL respectively and presented arguments for the validity, reliability, positive impact and practical nature of the test their institutions produce. Then, in turn, the discussants, Alan Davies (Edinburgh University), Amy Yamashiro (University of Michigan) and Liying Cheng (Queen's University) reviewed the tests in light of the ethical and technical standards expected of such high status, high-stakes and high volume testing instruments. The event served to demonstrate the commonality of many of the issues facing the three tests and a general comment from the discussants was that the three tests themselves acted as benchmarks for other large scale tests around the world.

Levels of spoken language ability: developing a descriptive common scale

Tony Green reported ongoing work on the Cambridge Common Scale for Speaking: an evolving frame of reference for the description of spoken language performance designed to provide test users with a clear explanation of the levels addressed by the various Cambridge tests.

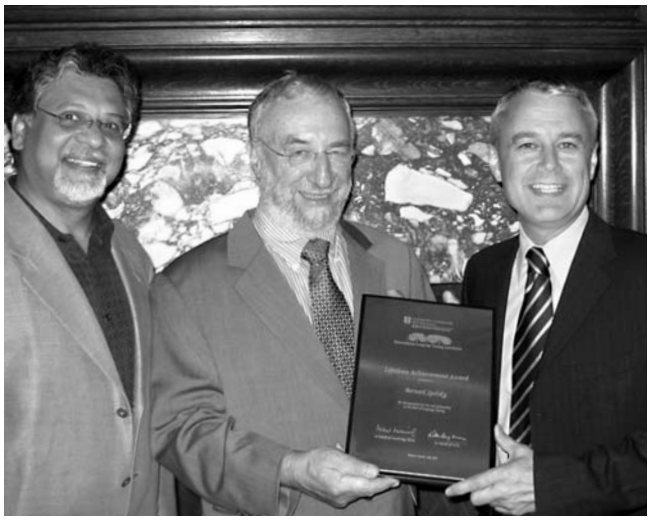
The current scale has been included in Main Suite examination handbooks since 1998, but needs to be updated to bring it into line with the recent development of the Common Scale for Writing and the expansion of the Cambridge ESOL examinations stable. Tony described how quantitative (test score and corpus) and qualitative (test discourse and participant protocol) analyses of test performance have been combined with insights from stakeholders to inform the ongoing validation and development process. He argued that the revised scale should reflect the Cambridge socio-cognitive test validation model (see the article by Shaw and Weir in *Research Notes* 21) and should reflect the interests of the range of stakeholders engaged in the processes of setting, taking or administering tests and of using test results.

Award presentations at LTRC 2005

During the gala dinner for LTRC 2005 which was held in Ottawa, the presentation of the UCLES/ILTA Lifetime Achievement Award was made to Professor Bernard Spolsky.

The presentation was made by Dan Douglas on behalf of ILTA and Nick Saville on behalf of Cambridge ESOL. Dan read out the citation (see ILTA online Newsletter or *Research Notes* 21) and Nick presented a plaque and a cheque for £500.

In making the presentation Nick drew particular attention to Bernard's contribution to the field of language testing in bringing an historical perspective to our current thinking. He recalled



Antony Kunnan, Bernard Spolsky and Nick Saville

Bernard's visit to Cambridge in the early 1990s when he was researching his book *Measured Words*. At that time he made use of the archives held in the basement of the Hills Road office and included references in the book and subsequently published articles based on what he uncovered – such as the significant but largely unknown work of Jack Roach during the 1940s. A legacy of this has been a concerted attempt by Cambridge ESOL since that time to document our work and to leave behind documents for future generations to review and evaluate. In particular the contributions of key individuals such as Jack Roach in his time, and more recently Dr Peter Hargreaves, need to be recorded and acknowledged appropriately – after all, while principles and concepts are essential, it is the people that make the difference!

To mark the occasion, Antony Kunnan and Nick Saville conducted an interview with Professor Spolsky during the conference; it is planned that this will appear as a feature in one issue of *Language Assessment Quarterly* in 2006.

Also at LTRC, Dr Annie Brown was awarded the 2005 Jacqueline A. Ross Dissertation Award by Educational Testing Service. The award recognises dissertation research that makes a significant and original contribution to knowledge about and/or the use and development of second or foreign language tests and testing.

Annie's dissertation, entitled "Interviewer Variability in Oral Proficiency Interviews" was written for the University of Melbourne and was selected from the largest pool of applicants ever received for this award. Annie's PhD research focused on the IELTS Speaking Test module introduced in 1989. In her acceptance speech she acknowledged the support she had received from Cambridge ESOL and the other IELTS partners during completion of her study. Her research findings were instrumental in informing the design of the revised IELTS Speaking Test introduced in July 2001.

For further information about Annie's award winning dissertation, visit the ILTA website: www.iltaonline.com

Cambridge ESOL staff were involved in LTRC, leading a workshop and giving a number of presentations. These will be reported in more detail in a future issue of *Research Notes*.

Annie Brown, award recipient, with her parents

Other news

Conference call: LTRC 2006

The 28th Annual Language Testing Research Colloquium will take place between June 29 and July 1 2006 at the University of Melbourne, Australia. The theme is 'Language Testing and Globalisation – Asia-Pacific Perspectives'. Pre-conference workshops will be held on June 28 on 'Policy and assessment in schools' (Penny McKay and Angela Scarino) and 'ConQuest (Rasch test analysis software)' (Ray Adams and Margaret Wu). The Samuel Messick Memorial Lecture will be given by Professor Mark Wilson from University of California, Berkley.

Cambridge ESOL staff will be attending and presenting at this event. For more information visit the LTRC 2006 website: www.languages.unimelb.edu.au/ltrc2006

Cambridge Assessment

The UCLES Group – of which Cambridge ESOL is a part – has adopted a new name 'Cambridge Assessment'. Cambridge Assessment is one of the world's largest educational assessment agencies.

The adoption of the new name does not change the legal status of the group – which remains a department of the University of Cambridge – nor will it change the way Cambridge ESOL presents itself.

For more information on Cambridge Assessment, visit www.CambridgeAssessement.org.uk