UNIVERSITY *of* CAMBRIDGE
ESOL Examinations

Research Notes

Issue 43/January 2011

A quarterly publication reporting on research, test development and validation

Printed in the United Kingdom by Océ (UK) Ltd.

Research Notes

Contents

Editorial Notes

Welcome to issue 43 of *Research Notes*, our quarterly publication reporting on matters relating to research, test development and validation within Cambridge ESOL.

The use of technology in language testing dates back to 1985 when the Language Testing Research Colloquium (LTRC) chose this theme for its annual conference. The very first *Research Notes* issue, published in 2000, contained an article on the use of computers in the Local Item Banking System at Cambridge ESOL. The theme of the use of technology was continued in issues 12 (2003) and 23 (2006) which addressed the relationship between technology and language assessment within Cambridge ESOL examinations. This issue of *Research Notes* is dedicated to the latest developments in technology harnessed for the purposes of language assessment at Cambridge ESOL.

The opening article by Sharon Jordan, Glyn Hughes and Cris Betts provides a broad overview of the use of information technology at Cambridge ESOL, discussing the associated benefits, issues and practices. The following three papers discuss the technological systems which support and facilitate the work of the external professionals who work on Cambridge ESOL examinations. Chris Hubbard outlines the development and impact of the Cambridge ESOL Professional Support Network (PSN), currently used for the co-ordination and standardisation of Speaking Examiners. Juliet Wilson and Murat Velioglu's article on Connect, a system through which computer-based tests are run, shares with us the benefits of the system from the perspective of test centres. Margaret Cooze's article on scoris® discusses the advantages of onscreen marking as well as the issues considered and addressed. The remaining papers are concerned with validation activities of computer-based (CB) and computer-adaptive tests from a variety of perspectives. Picking up a thread from Cooze's paper, Ardeshir Geranpayeh investigates onscreen marking via scoris® from a different angle, examining the comparability between onscreen and paper-based marking. Lucy Chambers and Kate Ingham outline the aspects of development of the *BULATS Online Speaking Test*, focussing on a proof-of-concept trial and alignment to the Common European Framework of Reference (CEFR). Following that, Lucy Chambers investigates the composition and revision strategies of a cohort of candidates who took the *Business English Certificate (BEC) Vantage* Test of Writing in the CB mode. Using an innovative method of data collection – the capture of the writing process through snapshots over the course of the test – Lucy explores the extent to which the assets of the CB medium are utilised during writing and how composition and revision strategies relate to the writing score achieved. Last but not least, Laura Cope and Andrew Somers discuss the challenges of paper-based pretesting and show how these challenges are addressed through a system of online pretesting in the context of computer-adaptive testing.

We finish this issue by reporting on the conference season and events Cambridge ESOL has supported. Ardeshir Geranpayeh reports on the *HR Magazine* conference (Hong Kong, July 2010), while Evelina Galaczi briefs us on the BAAL TEA SIG conference (Nottingham, November 2010). Angeliki Salamoura and Martin Nuttall report on the English Profile Project and ALTE events respectively. Finally, Lynda Taylor provides a brief on the latest volume in the SiLT series.

Technology in assessment

SHARON JORDAN ASSESSMENT AND OPERATIONS GROUP, CAMBRIDGE ESOL
GLYN HUGHES ASSESSMENT AND OPERATIONS GROUP, CAMBRIDGE ESOL
CRIS BETTS ASSESSMENT AND OPERATIONS GROUP, CAMBRIDGE ESOL

Introduction

Technology in assessment and Cambridge ESOL's application of developments in technology are themes that have recurred many times over the years in *Research Notes*. Issue 1 of *Research Notes* in March 2000 contained an article showcasing a newly developed software application for the hosting and maintenance of all Cambridge ESOL test material and metadata: Local Item Banking System (LIBS) (Beeston 2000:5). The main benefits of this new technology put forward in 2000 mirror the benefits we continue to aim for today via technology: greater efficiency, improved and harmonised processes and ability to meet the changing needs of those who use our assessment services.

Information Technology has moved on rapidly since March 2000, and indeed has come a long way since the last issue of *Research Notes* dedicated to technology in May 2003. Vast improvements in internet connectivity worldwide in addition to increases in connection speeds have made it possible to deliver a significantly broader range of content over the internet. This, combined with the increasing processing power of the modern computer and its increasing prevalence in offices, schools, colleges and indeed homes, has made this broad range of content accessible to an ever wider audience.

Throughout this time, Cambridge ESOL's use of technology in assessment has also expanded significantly, from the growth of computer-based testing to the introduction of onscreen marking, online results and online results verification. Although the use of technology in our assessment services has increased dramatically throughout this period, the principles underlying them have remained the same. Cambridge ESOL's approach has focused on:

- increasing choice
- improving the service we provide to centres and candidates
- maintaining and improving the quality of our exams.

This article provides an overview of how these principles are reflected across the developments that have occurred. This theme is then taken up in more detail throughout this edition of *Research Notes*.

Computer-based testing

From a candidate's perspective, the most obvious application of technology to assessment is computer-based (CB) testing. Technology applied to testing in the form of offering candidates tests via computer aims to make assessment more appealing, efficient and serviceable (Chalhoub-Deville 2001) and brings with it many advantages, including a positive impact on those who enjoy using computers, ease of administration, speed and reliability of marking, greater security, greater motivation to some candidates and arguably a more friendly interface than paper-based tests.

Cambridge ESOL has been producing computer-based tests since the mid 1990s, when *Communicat* and CB *BULATS (Business Language Testing Service)*, both innovative computer-adaptive language tests using an award-winning algorithm, were first introduced on CD-ROM. Ten years later, computer-based *BULATS* is one of Cambridge ESOL's most popular tests and is now taken online, on demand worldwide via a high specification online test administration and delivery system. This test administration and delivery system has also been extended to offer other assessment services such as online placement tests. Advances in technology have made life easier for test administrators who now simply need to log on to the testing system to run tests 24 hours a day, 365 days a year. There is no longer a need to order tests in advance of the test date or for installation of CD-ROMs or other software, and test results are available immediately. Cambridge Assessment's specially developed computer-based test delivery system 'Cambridge Connect' was launched in 2005 with a computer-based version of *Cambridge English: Preliminary (PET)*. Cambridge Connect interfaces with Cambridge Assessment back-end systems to allow a smooth end-to-end testing process from test creation on LIBS to test delivery to results availablity on ESOL Online. There are now 11 Cambridge ESOL examinations on Connect, with plans for more including *International English Language Testing System (IELTS), Cambridge English: First (FCE) for Schools* and *Cambridge English: Proficiency (CPE)* to be added in the near future. Technology in the form of CB testing is giving candidates more choice in terms of the test delivery mode but also in terms of when they take tests, with CB testing allowing Cambridge ESOL to now offer almost 700 test sessions a year. Supplementing paper-based (PB) assessments with computer-based versions has also enabled Cambridge ESOL to improve the service we provide by significantly reducing turnaround times. Candidates can enter for CB tests as little as one week before the test date and get their results just two weeks after they have sat their test.

Cambridge ESOL's approach to CB testing has always been to use CB platforms to increase the choice available to candidates and to provide tests that are fit for purpose: not to replace existing paper-based tests but to continue to offer paper-based versions of tests while also giving candidates the opportunity to take the test on computer if this is the candidate's preference (Blackhurst 2005, Jones &

Maycock 2007). All of Cambridge ESOL's CB tests on Cambridge Connect are computer-based equivalents of existing paper-based tests. The format of the test is the same on computer as on paper; it is simply the delivery method that is different. A major driver for having CB tests as additions to paper-based (PB) tests rather than replacements for them is the notion of bias for best (Jones 2003:4). In the early days of computer-based testing, there were concerns that construct irrelevant variance may be introduced as a result of familiarity with computers becoming a feature of what is being tested (Huff & Sireci 2001). However, as familiarity with computing has increased, concerns have also focused on whether handwritten responses are the most appropriate way to test writing (Russell & Haney 2000). Research into this field continues, as evidenced by Chambers' article (Chambers 2011 in this issue), which focuses on whether candidates make use of the potential benefits offered by composing written texts on a computer. By producing parallel CB and PB tests, we give candidates the opportunity to select the medium in which they feel most comfortable.

However, having two different test delivery modes for the same test on offer simultaneously, where the results from both modes are presented as comparable, brings its own issues and has necessitated continued research into the comparability of PB and CB testing. There has been extensive research in this area; see for example, Choi, Sung Kim & Boo (2003) and Taylor, Jamieson, Eignor & Kirsch (1998), as well as studies on paper-based and computer-based *IELTS* (Blackhurst 2005, Green 2004, Green & Maycock 2004, Maycock 2004, Maycock & Green 2005, Shaw, Jones & Flux 2001, Thighe, Jones & Geranpayeh 2001), computer-based *BULATS* (Jones 2000) and computer-based *PET* (Hackett 2005:12). As it stands, research generally concludes that test results are equivalent across test modes, and Cambridge ESOL's position is that candidates' interests are best served by offering them choice: 'Generally, we should agree that fairness will be best served by ensuring that candidates can choose the test format which they believe allows them to demonstrate their ability to the full – a "bias for best" approach' (Jones & Maycock 2007:12).

Indeed since these studies were carried out, computer systems have progressed and people have become much more used to dealing with text onscreen, so it is reasonable to assume that some of the extraneous variables which may have existed in these studies (e.g. quality of onscreen text or participant familiarity with computers) should be less of a factor today, and hence there is greater likelihood that test scores across PB and CB modes of tests are even more comparable now than in the past.

In terms of test delivery, Cambridge ESOL has taken different approaches depending on the precise nature of the tests involved. The exams delivered on Connect, including *Business English Certificates (BEC)*, *Cambridge English: Key (KET)*, *Cambridge English: Preliminary (PET)*, *Cambridge English: First (FCE)* and *Cambridge English: Advanced (CAE)*, are linear tests. These tests aim to build up a detailed picture of candidates within a relatively narrow ability range. As such, they are well suited to rigorous task-based assessment that tests a broad range of sub-skills at a given level. In addition, the high-stakes nature of these tests means rigorous security features are essential for running tests and tests must run with a very high degree of reliability. This means that, although Connect tests are delivered to test centres via the internet, they are administered to candidates offline via a centre's Local Area Network (LAN) and using software installed on local computers. This gives the increased confidence and increased security required for higher-stakes testing. Wilson & Velioglu (2011 in this issue) report in more detail on experiences of test centres using Connect.

With *BULATS*, a different approach was taken. In 2009, an online version of CB *BULATS* was launched. Unlike the tests delivered on Connect, *BULATS* tests are delivered fully online via a web browser. The major advantage of this approach is that it requires no software to be installed. This is an important factor for *BULATS* as tests are taken in a wide range of locations, including company premises. This delivery method also works very well for computer-adaptive tests such as *BULATS* as a large item bank can be hosted and maintained at Cambridge Assessment with items drawn down from the bank in real time as candidates take their tests.

With the benefits of technology also come the drawbacks and, as Bachman (2000:9) points out: 'The challenge in applying (such) technologies to language assessment will be to recognize not only the potential benefits but also the limitations of these technologies.'

One of the great challenges of an online test is the internet itself and its variability worldwide. Although a solid system can be built, it is also essential to consider issues such as bandwidth, latency, firewalls, PC configurations, etc. Whilst some of these cannot be controlled, technology itself can enable greater information and automated checks before tests are administered. Both *BULATS Online* and Connect systems continue to explore and use new technologies to enhance these services.

The coming year will see a major technological innovation for Cambridge ESOL with the impending launch of an online Speaking test, a first for Cambridge Assessment. Chambers & Ingham (2011 in this issue) detail the extensive trialling that went into producing the *BULATS Online Speaking Test*. Although, by its nature, a computer-mediated speaking test focuses a narrower range of the speaking construct than a face-to-face test (Galaczi 2010), the *BULATS Online Speaking Test* is an example of Cambridge ESOL's increasing focus on extending the range and availability of fit-for-purpose tests. In this instance, such a test is able to deliver the information that customers want in a quick and easy manner: companies worldwide are interested in an easy way of administering a Speaking test that will give an indication of overall speaking ability.

Test production and processing

The expansion in the range and number of tests we offer has been dependent on developments in the use of technology when producing and processing tests. The use of LIBS over a period of more than 10 years has

enabled Cambridge ESOL to systematically track and review task and item metadata (Marshall 2006:3). As mentioned by Jones (2003:3), the use of item banking has led to increasing applications of latent-trait theory to language testing. Swift turnaround times for *IELTS* and for CB tests are only possible when we have precise information regarding the level of ability the items in our tests measure.

Computer-adaptive testing (CAT) is predicated on the application of latent-trait theory to language testing and is a good example of how we have been able to make use of technology to increase the sophistication of our measurement processes. This is a process that started with LIBS, continued with the development of the CAT algorithm for *BULATS* (Maycock 2007) described above and is a continuous process. The introduction of SAS software has in the past 12–18 months enabled us to make huge improvements in the way we handle data. The use of SAS makes it possible to automate the creation of data files and control files to use external packages in a seamless process, minimising the opportunities for human error in data manipulation. Beyond this we now also have the tools to be able to carry out further analyses on a more frequent basis. For instance we have enhanced malpractice analysis by having the power to compare candidates' response patterns with all other candidates. Future plans include enhancements to test development, new analysis routines to provide further evidence on the functioning of test items, and further tests based on the latent-trait model. The improvements in data handling outlined above will almost certainly continue to influence the processes by which we construct tests and grade performance.

Any increase in the number of tests Cambridge ESOL produces is heavily dependent on the ability to pretest potential test material before live use. As paper-based pretesting has some limitations in terms of how efficiently and speedily material can be pretested, we have been looking to technology to find alternative and complementary solutions. A recent development in 2010 has been the implementation of online pretesting via our online computer-adaptive tests. Cope & Somers (2011 in this issue) outline the significant benefits that technology is bringing to pretesting and test production.

Automarking

Hand in hand with the increasingly sophisticated use of statistical data, we are also using technology to improve the rigour of marking procedures. In September 2010, we launched a new automarker which is used for all Cambridge ESOL computer-based tests. The automarker marks all short-text responses. When generating keys, the automarker references predetermined algorithms for date, time and currency keys. It also references a dictionary feature that contains alternative spellings and acceptable misspellings where appropriate. Test producers are able to configure a number of options depending on the testing product, including whether to accept misspellings or whether the automarker should be punctuation sensitive. The use of the new automarker has enabled us to ensure consistency of approach, where desirable, across all of our examinations.

Examiner management

Cambridge ESOL manages quality assurance processes for approximately 20,000 Speaking and Writing test Examiners worldwide. These processes are designed to ensure that all Cambridge ESOL Examiners are equipped to deliver fair and consistent assessments in live test environments and typically take the form of practice marking of non-live candidate performances via filmed Speaking tests or copies of Writing scripts. Since 2008, the majority of these activities for Speaking Examiners have been delivered via the Professional Support Network (PSN), which is described in more detail by Hubbard (2011 in this issue). Plans to make use of these technological solutions for Writing Examiners are also well advanced.

Similarly, Cambridge ESOL has been able to take advantage of the increased use and improved capability of technology to support the live operational work of Writing Examiners. The onscreen marking package *scoris assessor*® has been in use by ESOL Examiners since 2008 (see Cooze 2011 in this issue). By March 2011, all examining for *Main Suite* and *BEC* tests will be supported by scoris®. The system allows examiners to access and mark scanned and anonymous candidates' scripts by logging into a secure website. It is acknowledged that examiners rating text produced in tests of Writing could potentially introduce construct-irrelevant variance in test scores (Schaefer 2008:467). However, the scoris® package provides a number of major enhancements to examiner quality assurance, as identified by Harding & Raikes (2002:6). Salient among these is that examiners' marking tendencies can be monitored more comprehensively and effectively online, and that this can be accomplished at an early stage in marking cycles, allowing senior examiners to intervene in the marking process if necessary. Another feature of scoris® software allows Cambridge ESOL to feed in 'gold standard' scripts (Raikes & Shaw 2005:7), where examiners blind-mark scripts which have been previously rated by a group of senior examiners. This allows a reliable comparison to be made and re-marking to be triggered where the individual examiner's mark is out of tolerance with the expert group, and also facilitates post-test statistical analysis of the sort more commonly associated with objective testing than performance testing (Shaw and Weir 2007:312).

An obvious quality assurance concern arising from the introduction of scoris® was that the onscreen marking mode might have an impact on the marks returned. Geranpayeh (2011 in this issue), describing research carried out by Cambridge ESOL, indicates that this is not the case.

Cambridge ESOL is currently developing a bespoke examiner management system called *Cambridge ESOL Online – Examiners* which, as well as delivering numerous administrative services (such as automatic capture of examiner availability and electronic invitations), will support enhanced standardisation across all examiner QA processes and ensure that the most reliable examiners are targeted for live marking. As ESOL increases the range and availability of its products, the need to engage English language teaching professionals with excellent credentials also increases, and one objective of the system is to improve the quality of the examiner experience by, for

example, providing examiners with a one-stop portal through which they can manage their interactions with Cambridge ESOL.

Making life easier

The processes outlined above have enabled us to reduce significantly the time it takes to issue results without compromising on the necessary quality assurance steps. We have also been able to use technology in other ways to make life easier for centres and candidates. Cambridge ESOL Online, launched in February 2009, enables centres to make entries, timetable tests, print statements of entry and access results, all in one online portal. A similar portal is also available to preparation centres. Finally, the results verification website has added greatly to the value of our certificates by enabling any stakeholder to verify with Cambridge ESOL the result obtained by a candidate. This gives reassurance to institutions that accept Cambridge ESOL qualifications, and also to candidates.

Advances in technology have also enabled us to provide better support for our stakeholders. In addition to the web services offered to centres, we also have an online Teacher Support website (https://www.teachers.CambridgeESOL.org/ts/) with lesson ideas, online practice tests and online courses. Improvements in technology are enabling us to build a community of practitioners that use our assessments and courses and to share best practice among them.

The future

Cambridge ESOL continues to invest in technology in order to make processes more efficient and reliable and to improve the range and quality of the services we offer. Future plans include the modernisation of our processing systems to allow for more finely grained reporting of results, based on, for example, sub-skills tested within a single test. We are also investing in new materials management solutions that we anticipate may revolutionise test production in the same way as LIBS did 12 years ago, making it possible to deliver even more tests within even shorter timeframes.

In the area of computer-based testing, as well as new tests, future developments include a new and more technologically advanced version of our Connect test delivery system, to be launched in 2011. This will be a more intuitive, robust and user-friendly system and will be flexible enough to accommodate new types of CB tests and new methods of delivery in the future.

We anticipate that demand for CB testing, although currently considerably smaller than demand for paper-based tests, will continue to grow at an increased rate, in particular as our candidature becomes increasingly made up of a new technology-savvy generation who will want and expect to take tests on computer or via other electronic devices. As this happens and as new technologies and media emerge and advance, we will be able to explore new and exciting testing methods and modes of delivery. Some areas under consideration, which we would hope to be able to report on in the future issues of *Research Notes*, are technologies used by the gaming industry, Voice over Internet Protocol (VOIP) technology and mobile phone technology.

Of course, by then, technology will have moved yet further which is why our aim must be to continue exploring what technology may have to offer the field of assessment. Cambridge ESOL will continue not to be led by technology, but to make best use of it, in order to enhance the assessment products and services we provide so that they best meet the needs of those who use them.

References

Bachman, L F (2000) Modern language testing at the turn of the century: assuring that what we count counts, *Language Testing* 17 (1), 1–42.

Beeston, S (2000) The EFL Local Item Banking System, *Research Notes* 1, 5–6.

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

Chalhoub-Deville, M (2001) Language testing and technology: past and future, *Language Learning and Technology* 5 (2), 95–98.

Chambers, L (2011) Composition and revision in computer-based written assessment, *Research Notes* 43, 25–32.

Chambers, L and Ingham, K (2011) The BULATS Online Speaking Test, *Research Notes* 43, 21–25.

Choi, I, Sung Kim, K and Boo, J (2003) Comparability of a paper-based language test and a computer-based language test, *Language Testing* 20 (3), 296–320.

Cooze, M (2011) Assessing Writing tests on scoris®: The introduction of online marking, *Research Notes* 43, 12–15.

Cope, L and Somers, A (2011) Effective pretesting: An online solution, *Research Notes* 43, 32–35.

Galaczi, E D (2010) Face-to-face and computer-based assessment of speaking: challenges and opportunities, in Araújo, L (Ed.) *Computer-based Assessment (CBA) of Foreign Language Speaking Skills,* Luxembourg: Publications Office of the European Union, 29–52.

Geranpayeh, A (2011) The impact of online marking on examiners' behaviour, *Research Notes* 43, 15–21.

Green, A (2004) *Comparison of computer and paper-based versions of IELTS Writing: a further investigation of Trial A data*, unpublished report.

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

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

Harding, R and Raikes, N (2002) *ICT in Assessment and Learning: The Evolving Role of an External Examinations Board*, paper presented at Second SCROLLA Symposium, Herriot-Watt University, 6 February 2002.

Hubbard, C (2011) Cambridge ESOL Professional Support Network Extranet: Development and impact, *Research Notes* 43, 6–10.

Huff, K L and Sireci, S G (2001) Validity issues in computer-based testing, *Educational Measurement: Issues and Practice*, 16–25.

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

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

Jones, N and Maycock, L (2007) The comparability of computer-based and paper-based tests: goals, approaches, and a review of research, *Research Notes* 27, 11–14.

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

Maycock, L (2004) *CB IELTS: A report on the findings of Trial A (Live trial 2003/4)*, unpublished report.

Maycock, L (2007) Using simulation to inform item bank construction for the BULATS computer adaptive test, *Research Notes* 27, 7–10.

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

Raikes, N and Shaw, S (2005) *ESM Marking Models and Quality Control*, unpublished report.

Russell, M and Haney, W (2000) Bridging the gap between testing and technology in schools, *Education Policy Analysis Archives* 8 (19), 1–10.

Schaefer, E (2008) Rater bias patterns in an EFL writing assessment, *Language Testing* 25 (4), 465–483.

Shaw, S, Jones N and Flux, T (2001) *CB IELTS – A comparison of computer-based and paper-based versions*, unpublished report.

Shaw, S and Weir, C (2007) *Examining Writing in a Second Language*, Studies in Language Testing volume 26, Cambridge: UCLES/Cambridge University Press.

Taylor, C, Jamieson, J, Eignor, D and Kirsch, I (1998) *The relationship between computer familiarity and performance on computer-based TOEFL test tasks*, TOEFL Research Report 61, Princeton, NJ: Educational Testing Service.

Thighe, D, Jones, N and Geranpayeh, A (2001) *IELTS PB and CB equivalence: A comparison of equated versions of the Reading and Listening components of PB IELTS in relation to CB IELTS*, unpublished report.

Wilson, J and Velioglu, M (2011) Using Connect: The test centre perspective, *Research Notes* 43, 10–12.

Cambridge ESOL Professional Support Network Extranet: Development and impact

CHRIS HUBBARD ASSESSMENT AND OPERATIONS GROUP, CAMBRIDGE ESOL

Introduction

In order to successfully produce and deliver a large range of examinations to candidates worldwide, Cambridge ESOL relies on an extensive cadre of professionals worldwide. Cambridge ESOL draws on a network of people trained in specific responsibilities and subject to ongoing Quality Assurance (QA) procedures in order to ensure continued high standards of work. As the network has grown and diversified so have the requirements for the ways in which Cambridge ESOL supports them, both as groups and as individuals. In late 2008, after six months of trialling and system development, Cambridge ESOL launched a web-based extranet system called the Professional Support Network (PSN), with the specific intention of enhancing and extending the support given to external groups.

The first group targeted in full was the Speaking Examiner cadre. This article will overview the process in terms of the development, the uptake and the impact of PSN, with a view to assessing the success of the system to date and identifying future stages to be followed.

PSN development

The ongoing support of Speaking Examiners is a well established system which includes annual standardisation of examiners, both in procedures and in marking. Prior to the launch of PSN, this stage of the quality assurance process, known as co-ordination, used to be completed in face-to-face meetings which required examiners to attend a series of meetings to cover each exam that they were eligible to examine for. Although this was a robust and proven system, it also had some limitations:

- it was time consuming for examiners who had to attend multiple meetings to cover a range of exams
- it was limited to a once a year focus, although examining sessions take place throughout the year
- session content and speed were dictated by the session programme and timing.

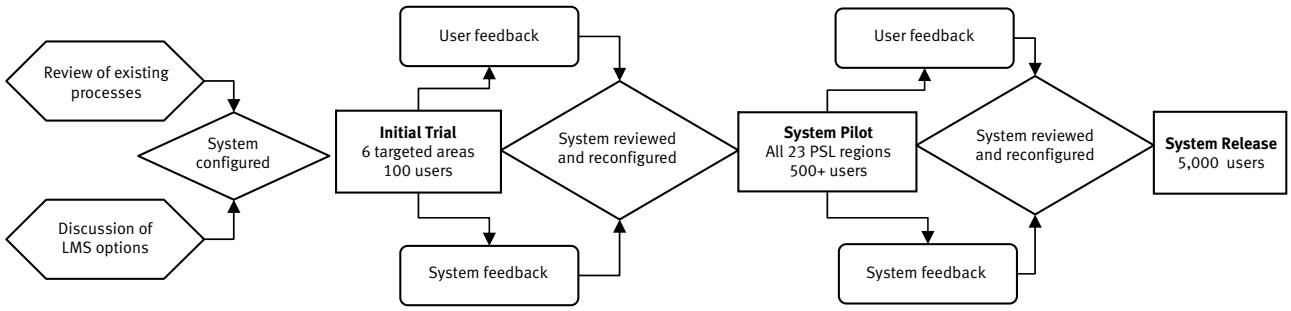
In response to the highlighting of these limitations by examiners and Centre Exams Managers, the development of PSN was initiated. This development sought to address these limitations by complementing an annual face-to-face meeting with online materials and tasks to improve the process and provide the following opportunities (Mitchell Crow & Hubbard 2006):

- access – allowing 24/7 access, with real time feedback and instant access to materials updates
- flexibility – allowing examiners to work with materials at their own pace and to participate in development tasks when it best suits them in the examining year
- autonomy – offering development activities in this mode offers examiners a certain level of freedom in taking responsibility for their own professional development.

The development process was designed to take best advantage of the knowledge and experience of those involved in the existing QA processes, and to combine them with the opportunities being offered by the introduction of a Learning Management System. Figure 1 shows the linear process adopted and exemplifies how the three main development stages were combined with consultation and analysis of outcomes in order to inform the following steps and the subsequent configuration of the system.

Space limitations do not permit a detailed look at each

Figure 1: PSN Development Processes and Stages



stage and the decisions taken. Instead, an overview will be given of the main outcomes and the focus of user feedback from the Trial stage and a brief overview of the Pilot stage. The aims of the trial were important because they sought to put in place and test a number of the founding principles for the organisation and configuration of the system, while the pilot was a full system 'user testing' environment that would give users and administrators access to materials and data in a live operational situation, and so verify the effectiveness of the earlier design decisions.

PSN trial

The Speaking Examiner hierarchy is divided into 23 regions, with each region made up of the exam centres within the country or group of countries designated to that region. Teams of Speaking Examiners (SEs) are managed locally by one or two Team Leaders (TLs), and TLs are in turn managed within the region by the Professional Support Leader (PSL) who may be assisted by one or more Regional Team Leaders (RTLs). In order to represent the varied regional scenarios that examiner teams operate in, and to ensure that user feedback would address relevance and usability of the system both in terms of the effectiveness of QA processes and the technical practicalities of accessing materials, a targeted group of participants was selected to take part in the initial trial.

Participants were invited from six of the Speaking Examiner regions, and involved PSLs, RTLs, TLs and SEs in each region. There were over 100 participants. Sample Speaking tests for *Cambridge English: First (FCE)* and *Business English Certificates (BEC) Preliminary* were included and the materials for each examination were divided into three stages which related to the way they would be presented in a face-to-face co-ordination meeting:

- Setting the Standard – a benchmarking exercise before being asked to assess
- Applying the Standard – an opportunity to practise applying assessment criteria with feedback provided
- Marks Collection – a once only submission to check examiners are marking within acceptable limits.

The fundamental objectives of the trial, and to a large degree of the eventual completed PSN system, were functionality based and are detailed in Table 1 from SE, TL and system perspectives.

In terms of achieving these functionality objectives, the trial was successful in fully or partially achieving all the 15 objectives outlined above. For those objectives which were only partially achieved, such as accessing video samples and other materials, the common and main contributing factor preventing their full achievement was system connectivity. This was mostly related to delivery of large video samples across a variety of end user system configurations and hardware capacities, and a major outcome from the trial was a focus on revising materials in order to present them in the most accessible format possible whilst still retaining a high enough quality to ensure that examiner assessments would not be adversely affected.

Trial participants were asked to complete a survey related to system usability and the outcomes of that survey are summarised in Table 2. Participants were asked to rate each of the questions on a scale from 1 (unsatisfactory) to 5

Table 1: Objectives of PSN trial (February 2008)

Speaking Examiner functionality	Team Leader functionality	System
1. Can watch sample video tests	1–6. Can complete all SE tasks	11. Easy to use – user guides
2. Can submit assessment marks	**In addition:** 7. Can enrol SEs into exam areas based on eligibility, when required	12. Successfully completes all SE and TL functionality
3. Can receive outcome and access support commentaries	8. Can check outcomes of SE work	13. Can present multi-media content
4. Can submit secure Marks Collection marks	9. Can access individual tests if required	14. Records user activity and outcomes
5. Can follow SE user guides	10. Can follow TL user guides	15. Operates without heavy support requirement
6. Can submit feedback		

Table 2: PSN trial: Feedback question scores

Question	Average score on a scale 1–5
Login and general access to the system	4.2
Navigation within PSN	4.2
The speed of operation on your computer	3.9
Video and sound quality	3.8
Instructions and user guides	3.8
Your overall impression of the PSN	4.1

(excellent), and were also given the opportunity to submit free text comments. The average scores for all six questions were all acceptably high.

The comments gathered at this stage were extremely informative. Many of the positive comments picked up on some of the main aims of the system:

> 'An excellent system which will save time and money [compared to] attending standardisation meetings.'
>
> 'Very helpful for finding the benchmark at the beginning of an examining session.'
>
> 'I liked it. Nice to be able to re-watch videos and read comments about candidates' performances to get a real feel for the exams and levels.'
>
> 'This could work well as an addition to face-to-face co-ordination as SEs often want a chance for supplementary standardisation.'

However, a number of users also identified areas that required focus and improvement, such as the accessibility of multimedia files:

> 'I can see that this could work, but I found it frustrating working on a narrow bandwidth.'

Other very useful comments related to the clarity of user instructions. The system was to be rolled out to examiner groups across the world without direct user training and so would rely on intuitive and comprehensible functionality and clear user guides. The following comment suggested there was still work to be done in this area:

> 'I was a bit disconcerted at first because I didn't have access to all areas. It may have been presented clearly enough for most people, but for some (including me) it may have to be spelled out even more simply.'

There was also feedback related to some of the extra administrative functionality provided to TLs, functionality that reflected procedures directly transferred from the existing QA process. Representative of the concerns raised in most comments is the following:

> 'Enrolling your SEs (to restrict their access) was not easy … this was time consuming.'

These comments highlighted the need for the system development process to include the understanding that moving existing procedures into an online environment requires an extensive review of the procedures themselves, and the principles that underpin them, so that the new environment would not be encumbered by constraints which result from the transfer of a physical process into an online environment. The subsequent review of the trial feedback focused on the following crucial aspects of QA procedures, and fed directly into the system reconfiguration for the pilot and live release stages:

- Restricted versus open access. The trial reflected the existing restricted access model. Prior to PSN, examiners only attended face-to-face meetings covering exams for which they were eligible to examine. Feedback from TLs and system administrators questioned the limiting of access in the online environment. It was subsequently agreed that the pilot would include an open access model with examiners required to complete tasks covering exams relevant to them, but being able to access materials relating to other exams if they wished to extend their knowledge and exposure to other assessment levels. Based on the fact that the Cambridge ESOL Speaking test assessment scales form a single continuous scale from A2 to C2 (on the Common European Framework of Reference) with overlap at the extremes of performance at each level, this was seen as a positive outcome.
- Security of reference marks versus instant feedback. One result of all co-ordination being covered in face-to-face meetings was that all reference marks and performance commentaries were held by the TL (the trainer), and referred to as required during the session. The access to marks and commentaries of some samples was restricted in the trial. The related user feedback expressed the need to show this information so that the experience could be more valuable. Instant feedback and access to information explaining potential differences in marks or confirming similarity in marks was seen as essential to allow examiners the autonomy to shape their own development.
- Local versus centralised administration. One aim of the trial had been to empower local TLs to populate and locally manage their examiner teams online. However, as identified above, some aspects of this were seen as being time consuming and outweighed benefits gained. Preferences were expressed for the centralisation of routine administration of the system, but with TLs requesting more ability to question and report on examiner outcomes at a local level, in order to best support individual examiners when needed.

PSN pilot and system release

The review and reconfiguration discussions that followed the trial concluded that the above QA procedural questions raised in feedback would be answered by giving examiners open access to all Speaking test areas of the system, providing instant feedback including marks and performance commentaries, and by managing routine administration centrally.

The pilot took place in May and June 2008 and included all PSLs, RTLs, and TLs from around the world; in total, it involved over 500 participants. The same objectives outlined in Table 1 were identified, with the small amendment to objective 7: 'TL can enrol SEs into exam areas based on eligibility' was amended to 'SEs are enrolled into exam areas and can access materials' and was moved into the system/central administration column.

During the pilot all functionality objectives were achieved, and all to an acceptable level of success. This meant that going live with the system would not result in large groups of examiners having to contact the PSN support team in order to carry out the basic functionality for which the system was intended.

Based on trial feedback, a description of minimum system specifications was circulated to TLs and SEs prior to the pilot and system release. In addition, the release

was supported by the production of existing DVD-based co-ordination materials that could be used in areas where access was problematic. Therefore, from system release, Cambridge ESOL predicted that approximately 50% of the SE cadre would move to the system immediately, and aimed for the uptake to increase from there to gradually include more and more of the examiner group.

PSN uptake

The PSN extranet has been operational for Speaking test Examiners since August 2008, and it formed an essential part of the process introducing a revised set of Speaking test assessment scales in the last quarter of that year. A measure of the successful introduction of a system such as this is uptake, reflected in the amount and rate of increase of users it is designed to support. The other major area to measure is the impact of the system on the elements of the QA process it set out to enhance. Both of these areas will be discussed here.

In terms of uptake, Table 3 outlines the usage since 2008. The data for 2008 includes the system usage during the trial and pilot stages (approximately 600 users).

Table 3: PSN usage figures (2008–2010)

	2008	2009	2010 (Jan–Oct)
Total logins	32,942	42,083	56,204
Individual users	5,079	6,820	7,780
Exam materials accessed	117,353	180,430	197,694
Marks Collections completed	10,001	16,193	18,937

From these figures we can draw the following conclusions in relation to system uptake and subsequent impact:

- On average, individual users logged on around six times each in 2008 and 2009, and seven times each in 2010. This increase in 2010 is encouraging and indicates sustained usage of the system as part of the overall QA process.
- On average, during each of the six or seven visits that they make each year, users accessed 20 exam related materials (e.g. sample Speaking test videos and support materials) and in 2009 and 2010 this rose to almost 30.
- Examiners are required to complete Marks Collection assessment tasks for each CEFR level that they examine at. In 2008, the average of 1.6 Marks Collection events were completed per user via PSN, and in 2009 and 2010 this rose to 2.7, again showing a confidence in, and acceptance of, the system.

PSN impact

The introduction of the PSN extranet system can be seen to have had the following impacts on the co-ordination stage of the QA procedures for SEs:

1) Reducing the number of meetings that examiners have to attend. All examiners now attend a single annual face-to-face co-ordination meeting.
2) Refocusing the content and format of those meetings to cover a better balance of procedural and assessment issues.
3) Providing examiners with flexibility to access and work with materials when they wish, and to do as little or as much of this as they find necessary.
4) Providing Cambridge ESOL with focused feedback on the standardisation materials we produce via system reports relating to proportions of marks awarded for every sample. This allows us to monitor whether particular exams, types of performance, levels of performance or assessment criteria, etc., pose more of a problem for examiners when they co-ordinate than others. If this is the case, we can ensure future performances and supporting commentaries target these areas and that we provide more practice opportunities.

Next areas of focus

The successful implementation and uptake of the PSN extranet to support the worldwide cadre of Speaking test Examiners has established a firm foundation on which to build extra applications and support for other groups of professionals. In terms of examiner groups the following aims have now been established:

- Speaking Examiners – The current aim is to support other sections of the QA 'life cycle', from initial recruitment through training to ongoing support and monitoring, via the PSN extranet.
- Writing Examiners – Cambridge ESOL intends to draw from, and build on, the success of online delivery of QA processes for Speaking Examiners, by extending this functionality to Writing Examiners (WEs). This project aims to deliver the same benefits achieved through online co-ordination for SEs, but with one important addition. Currently WEs are co-ordinated using scripts which are made available after the test is taken. Under current procedures, these are then marked by a panel of senior examiners, and then used to standardise the rest of the team (Shaw and Weir 2007:276). Moving to the Speaking test QA model of conducting the co-ordination process before the test date will facilitate the faster release of results to candidates, and provide important QA enhancements in terms of the reliability of reference marks used in standardisation.

The process and system outlined in this paper has been in operation for Speaking Examiners for two years and, as discussed above, has been deemed to be a successful addition to examiner QA processes. It can also be extended to a wide cadre of other professionals (e.g. Centre Inspectors, Teaching Awards providers), on whom Cambridge ESOL relies worldwide in order to support its operation. The outcomes described here in terms of system design and usage, and approaches to transferring existing procedures into the online PSN environment, can all be applied to reviews of the QA processes for other sections of the external cadre covering other responsibilities. This will allow Cambridge ESOL to also support these professionals in the most efficient, effective and flexible manner it can.

References

Mitchell Crow, C and Hubbard, C (2006) ESOL Professional Support Network Extranet, *Research Notes* 23, 6–7.

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

Using Connect: The test centre perspective

JULIET WILSON CUSTOMER SERVICES GROUP, CAMBRIDGE ESOL
MURAT VELIOGLU CUSTOMER SERVICES GROUP, CAMBRIDGE ESOL

Introduction

Over the last 12 months, there has been a huge uptake of computer-based testing across the Cambridge ESOL exam centre network. We now have over 350 centres running computer-based tests on Connect and at least 200 others who are currently being trained or completing their application process. In this article, we will look at the reasons for this increase in interest and describe a number of cases where centres have taken advantage of the many benefits which computer-based delivery offers to candidates and centre administrators.

Background

When computer-based tests were first launched in 2005, a number of Cambridge ESOL centres were initially reluctant to introduce them. Many believed that they did not have sufficiently up-to-date equipment or the requisite number of computers. Others said they did not have the technical expertise to administer the tests. Moreover, for centres who were used to filling huge exam halls with candidates taking paper-based tests, the logistics of offering a large number of tests on computer seemed complex and even intimidating. There were also concerns about the security of computer-based testing and the possibility of malpractice such as candidates being able to access helpful websites during the tests. Some teachers were resistant to, or even felt threatened by, the introduction of technology in their examination preparation classes. In some parts of the world, centres reported that there was no appetite in the market and candidates were not requesting to take their exams on computer, preferring to stay with the more traditional pen and paper method.

Examination centres embrace computer-based testing

Over the last year, there has been a huge increase in the number of centres embracing a computer-based way of administering language tests. Below we discuss the reasons behind this dramatic change:

- In 2011, there are 123 computer-based test dates and 566 opportunities for candidates to sit a computer-based test. This is a radical change from the situation a few years ago when candidates could only sit a paper-based Cambridge ESOL exam twice a year. This increase in the number of dates and the possibility of running more than one session on each test date means that centres are able to offer candidates more choice and flexibility about when they take the test. In this way, they are able to meet the needs of their candidates more effectively.
- There are now a very large range of Cambridge ESOL examinations available on Connect: *Cambridge English: Key (KET)*, *Cambridge English: Preliminary (PET)* and the respective 'for Schools' versions, *Cambridge English: First (FCE)*, *Cambridge English: Advanced (CAE)*, *Business English Certificates (BEC) Preliminary*, *Vantage and Higher, Skills for Life (SfL)* and *Teaching Knowledge Test (TKT)*.
- The reduced time for making entries and the quick turnaround of results suit candidates who need to take an exam quickly and receive their results within two weeks.
- Computer-based tests reduce a centre's administrative load as there is no storing, packing and return or secure destruction of question papers required. Centres can also spread their administrative work over the year by organising a number of smaller sessions rather than one or two very large sessions.
- Computer-based testing can also be cost-effective as centres can save on postage and delivery charges. Centres can also avoid the stress which can be caused if materials are held up in customs or delayed in transit.
- The interface of Connect is very user-friendly and candidates appreciate the benefits of online features such as the help function and the timer, and the fact that they can edit their answers onscreen. In addition, candidates take the Listening test using headphones so they can adjust the volume to suit their particular needs. All these features combine to ensure that the test takers can do their very best when they sit for a computer-based test.
- Online practice tests are available to help teachers and candidates prepare for the exams. Teachers have realised that it is not necessary for them to understand the technology to prepare candidates for success in the computer-based exams. The content of these exams is exactly the same as that of the paper-based exams and

the research carried out by Cambridge ESOL confirms that the mode of marking (paper-based vs. onscreen) has no impact on examiners' behaviour and the resulting marks (see Geranpayeh 2011 in this issue).

Becoming a computer-based exam centre

The Customer Services Group registers, trains and supports Cambridge ESOL exam centres and is responsible for ensuring that they run the examinations in line with Cambridge ESOL regulations. Over the last year, the procedure for becoming a computer-based centre has been made as easy and straightforward as possible, while maintaining its rigour. The latest procedure consists of the following steps:

1) Centres complete an application form, including confirmation that their computers meet the published technical requirements. (The Cambridge Connect software system uses a standard computer network with candidate workstations linked to an administrator workstation. Centres only need five computers to set up as a centre and although the technical support person needs to be computer literate, there is no need for them to be an IT specialist.)
2) A member of the Application Support team in Cambridge confirms the centre's technical suitability to run the exams.
3) The test administrator and the technical support staff complete a training session.
4) The centre staff install the software at the centre and run a dummy test.

Training and support

Cambridge ESOL Customer Services offers centres either online training via Moodle or face-to-face training which takes place over one or two days. If centres choose to do the training online, they must also complete a short test to confirm their understanding of the technical requirements and the exam day regulations. Many centres have chosen the face-to-face training and have appreciated the workshop and hands-on approach as well as the opportunity to meet other centres who are also beginning this new way of working. Since June 2009, 28 face-to-face sessions have taken place across the world in a number of destinations from Australia, Brazil and Canada to India, Spain and the UK. The training equips the test administrator and technical support staff to install and run the tests, including basic trouble-shooting.

However rigorous a training programme is, centres are understandably nervous as their first computer-based test day approaches. To help give centres the confidence and reassurance they need, the Cambridge ESOL Application Support team sends all centres a series of emails leading up to the test day to remind them of the key tasks they need to do in preparation for the test. On the day of the test, Application Support offers on-call 24/7 telephone and email support. The effort invested in training and support has resulted in success, as the Centre Exams Manager of QLS, Greece, testifies: 'The training that we were offered by our colleagues in Cambridge did pay off along with the extensive support we were given during the duration of the exam.'

Security and quality assurance

Cambridge ESOL takes the security and integrity of its tests extremely seriously and works closely with centres to ensure that the administration of examinations meets requirements and that the security of all confidential materials is maintained. The security of the examinations offered on Connect is protected by state-of-the-art encryption and an onscreen lock-down facility. Regular inspections at centres ensure compliance with all regulations. In addition to general criteria related to security of materials and conduct and supervision of the staff, an inspector visiting a computer-based exam session will also assess centres using particular criteria related to the technical equipment and support available. All this ensures that for the test taker, the exam experience is fair and positive, and for centres, the quality requirements stipulated by Cambridge ESOL are clear and transparent.

Feedback from test centres

A number of our test centres were asked to tell us about their experiences with Connect. Their feedback outlined the benefits which we had anticipated. The main themes which came through were: improved flexibility, enhanced customer service, easier administration and the up-to-date image which computer-based testing gives a centre.

The Cambridge ESOL examinations centre at Winterthur, Switzerland was one of the first centres to offer computer-based tests in 2005. Peter Kaithan, Director of Marketing and Communications, saw the need for a system that would allow his centre to run exams with more flexibility and with less administration. He says: 'Today, with the demand for computer-based testing on the rise, we are slowly but surely experiencing the return we hoped for. Even though a quite expensive undertaking, the benefits today, and particularly over time, will outweigh the cost. Beyond the financial aspect, providing exams on computer also gives us a progressive and advanced image as an examination provider which made the investment in a multimedia PC lab worthwhile, particularly since the infrastructure can be used for other purposes such as teacher seminars.'

Veronica Cameron, Centre Exams Manager at Buenos Aires Open Centre, started offering computer-based tests in October 2009 with the primary aim of offering a better service to her customers: 'We can give them more flexibility with enrolment, payment and they can see their results in two weeks' time. We rented a suitable venue located in the heart of our city. The desks were specially built for the computers and candidates can complete their tests in a quiet and comfortable environment. Most candidates have sent very positive feedback and many are willing to repeat the experience.' As well as pleasing her candidates, centre staff have also been very positive about the new way of

working: 'Not having to pack papers and boxes and the saving in postage and customs is greatly appreciated by supervisors and centre managers.'

Margaret Fowler, the British Council Country Examinations Manager in Italy reiterates the above-mentioned benefits. She reports that candidates have shown a growing preference for computer tests in Italy over the last two years: 'The word count facility is very popular and gives candidates more time to concentrate on actually writing their answer rather than totting up the number of words. Candidates like the fact that they can do the whole exam in one day including the Speaking test because the sessions are smaller.' For the centre, one additional benefit which has emerged is that they need to train up fewer Speaking Examiners: as there are more exam sessions spread over the year, there are smaller numbers of candidates per session. Margaret summarises the reasons for her involvement and commitment to computer-based testing: 'It has become a matter of principle for us at the British Council, Italy to be in the vanguard of whatever is new in testing.'

One of the newest recruits to computer-based testing is Harmon Hall in Mexico. The Centre Exams Manager, Rommel Mandujano, and his academic team, have been eagerly anticipating the launch of the new Cambridge ESOL computer-based testing system which will be delivered during 2011 and explains: 'The new system will offer Harmon Hall two main benefits over paper-based exams: firstly it will allow us to administer Cambridge ESOL examinations in a Mac environment and secondly we will be able to register entries and deliver results within a shorter period of time, which aligns with our short study terms of one to two months.'

Connect – the future

After each computer-based session, the Customer Services Group collects feedback on centres' experience with Connect. The aim of this is to ensure that the user experience remains positive. This feedback has been integral in helping us to define requirements for the 'next generation' of Connect. Some of the new features which are being introduced as a direct result of this feedback are:

- a much simpler installation process, requiring even less technical expertise at centres
- compatibility with Windows 7 and Macs
- the ability to run Connect on laptops, opening up many more possibilities for offering tests in different venues.

As the test centre perspective is so important, Cambridge ESOL is running a number of trials of the new system prior to the launch in 2011. By remaining focused on the test centre perspective, we can ensure that we are able to continue to deliver the many benefits of computer-based testing to centres and their candidates.

References

Geranpayeh, A (2011) The impact of online marking on examiners' behaviour, *Research Notes* 43, 15–21.

Assessing Writing tests on scoris®: The introduction of online marking

MARGARET COOZE ASSESSMENT AND OPERATIONS GROUP, CAMBRIDGE ESOL

Introduction

Assessment bodies have been taking advantage of technological developments in recent years in the form of computer-based testing. They have also been able to draw on the wealth of performance data provided by these technological developments to inform exam revisions and to feed into other research. Over the past five years in particular, Cambridge Assessment has widened the use of technology to include onscreen marking of extended essay responses. Onscreen marking is now in use for Cambridge ESOL computer-based tests (CBT) and paper-based tests (PBT). The present article provides background to the onscreen marking system being used, discusses its benefits and summarises the training which was provided to a key group of stakeholders: Writing Examiners of *Cambridge English: Advanced* (also known as *Certificate in Advanced English (CAE)*).

Background

As the candidature for Cambridge ESOL exams continues to grow, so does the requirement for more flexibility in the provision of exam dates and formats. In meeting this demand, it became evident that Cambridge ESOL would need to harness technology to ensure that it continues providing results to candidates accurately and promptly. As long ago as 1999, Cambridge Assessment started investigating a system whereby scripts (candidates' responses to Writing test questions) are scanned and the resulting digital images loaded to a web-based system for examiners to download and mark. Initial trials of onscreen marking were carried out in 2000 and 2001, but it was not until 2003 that research into onscreen marking using the scoris assessor® marking system and collaboration working with our technology partners RM plc. began in earnest. scoris® and the system it is part of, Electronic Script

Management (ESM), have already contributed considerably to improved Quality Assurance. They will be discussed in this article and possibilities they offer for new developments will be outlined in the last section.

Electronic Script Management (ESM)

Electronic Script Management (ESM) is one part of a strategic corporate programme which supports the online marking of candidate responses. It enables onscreen marking of candidates' responses, allowing for remote co-ordination and standardisation of examiners as well as acting as an electronic system for tracking scripts. RM's scoris assessor® package has formed a major part of this management system. It is an established software package for marking online, created as a result of RM's stated vision to improve the marking process for both assessment bodies and examiners: 'For individual markers, scoris® presents a secure online environment in which they can receive, view and mark exam scripts on their own home computers. For examination boards, it provides a web-delivered workflow system, which controls the process of allocating exam scripts to markers, collecting marks and monitoring the quality and consistency of markers' work' (RM plc.).

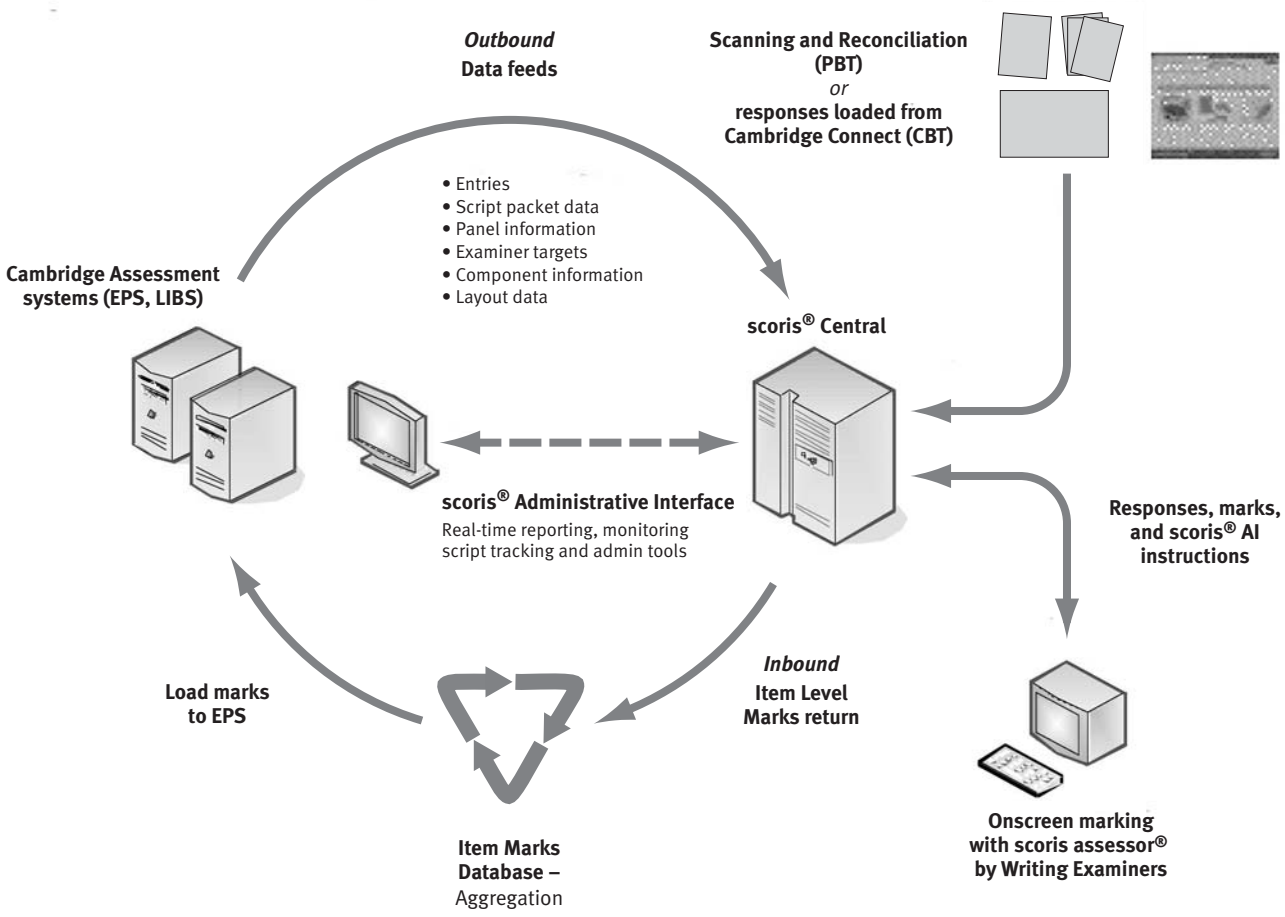
The ESM marking process, based on the use of scoris®, is illustrated in Diagram 1 and involves:

- the transfer of entry and Writing Examiner information in data feeds from Exams Processing System (EPS) to RM
- creating digital objects from candidate scanned responses in the case of PBT responses
- loading digital images from PB and CB tests onto a server (scoris® central)
- distributing the digital objects electronically so that examiners can log onto the server and mark onscreen at home
- capturing marks and examiner annotations from the marking process electronically in the Item Marks Data Base (IMDB)
- post-marking processing in the Cambridge Assessment Exams Processing System (EPS) and Local Item Banking System (LIBS)
- results issue and storage.

How scoris assessor® works

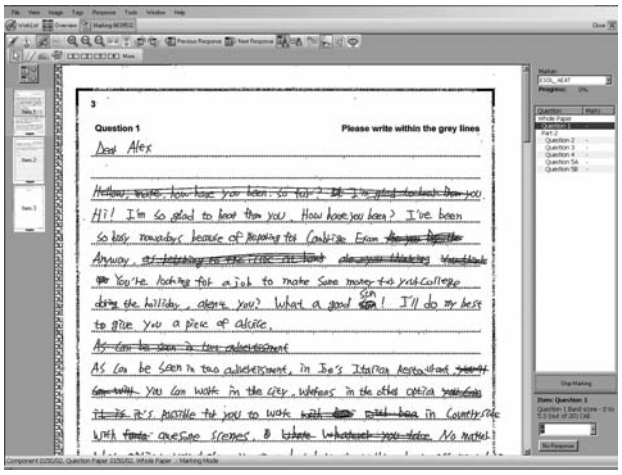
In order to mark on scoris®, examiners are first asked to check that they have the necessary PC specifications and broadband connection. Following that, they download the marking system and access the scoris® server with a secure login and password. The system provides examiners with tools to view responses, use zoom tools to facilitate reading, mark annotations electronically and note comments, as well as recording marks. Visuals 1 and 2 show the examiner view of CBT and PBT responses on scoris®.

Diagram 1: The ESM marking process using scoris®



Visual 1: Examiner view of CBT response on scoris®

Visual 2: Examiner view of PBT response on scoris®



Marks are stored and returned to the awarding body following a grace period during which examiners can return to scripts and adjust marks in light of feedback received from a supervisor.

The benefits of online marking

There are various benefits of online marking, for examiners and examining bodies alike. The immediate benefits for examiners include a reduction in administration and the removal of non-core tasks, such as dealing with script collection and postage. This allows examiners to focus on the professional aspect of marking and results in increased productivity. Operationally, the time spent dealing with tracking script packets is eliminated and overall administration is reduced at the awarding body. This also contributes to increased security of scripts as they are scanned centrally and stored on the secure server. In this way, PBT marking has been very much streamlined.

The design of the marking system contributes to improved quality assurance in a number of ways. Marking is anonymous as scripts are presented without the candidate and centre details identified and are randomly allocated to examiners rather than being presented as a batch of responses from a particular centre or country. This prevents examiners from becoming 'tuned in' to typical errors from speakers of particular language groups which may influence their marking, thus improving the fairness of the assessment.

As responses and marks are stored on the scoris® server, they are available for the awarding body, Team Leaders (TLs) and Principal Examiners (PEs) to view at any time once examiners have marked them. This transparency allows monitoring of examiners to take place easily with TLs and PEs able to identify any inconsistent marking very early in the marking window. This allows early intervention, with examiners being stopped from marking if necessary, to ensure that marking is carried out to a consistently high standard. In addition the frequent return of marks allows for early analysis of performance data in preparation for the grading process.

The history of scoris® in Cambridge ESOL

scoris® was first trialled and used for live marking of CBT Writing components during 2007 and 2008. The candidature for the first live sessions was relatively small and marking was carried out residentially in Cambridge. Following that, scoris® was introduced into the marking of paper-based components along with the launch of the updated *CAE* in December 2008. Prior to this point, *CAE* had also been marked by teams of examiners working residentially. The examiners are led by a Principal Examiner (PE) who is responsible for the overall marking of the paper. The PE manages a number of Team Leaders (TLs) who, in turn, manage teams of Writing Examiners.

Ahead of the first live session, in the summer of 2008, a representative group of *CAE* Principal Examiners, Team Leaders and Writing Examiners was identified to take part in a study comparing the effect of the mode of marking – paper-based and onscreen – on examiners' behaviour (see Geranpayeh 2011 in this issue for more detail). The opportunity was also taken to observe examiners marking online, to obtain soft feedback on the marking experience itself and to make use of the feedback to inform examiner training. Interviews with examiners investigated examiners' technological background, experience and ability along with their perceptions of onscreen marking. Research carried out by Cambridge Assessment in 2004 noted that 'research into examiners' experience and views about onscreen marking had mixed findings, and these should be explored fully as part of the development of any new system' (Raikes, Greatorex & Shaw 2004). Raikes et al (2004) focused on examiners in the other business streams of Cambridge Assessment (OCR and CIE), but the issues raised by examiners were clearly relevant to the ESOL scenario as examiners worked in very similar ways. Examiners were concerned as to whether marking onscreen would be as reliable as the traditional marking methods and whether the outcomes would be the same for candidates. A number of examiners felt that a high level of computer ability would be required to mark onscreen and that high-specification computers would be required. It is clear from this study that there was considerable examiner apprehension about the shift to onscreen marking. ESOL Writing Examiners showed

that similar concerns and anxieties existed to those reported by Raikes et al (2004). Based on this, it was decided that practical face-to-face scoris® training should be carried out with all examiners in order to make the transition from paper-based to onscreen marking as smooth as possible. It was also decided to present examiners with a rationale for the move to onscreen marking and the benefits which examiners, candidates and the awarding body would see from this shift. We were thus able to tailor training to the needs of examiners and to provide a suitable level of support for initial marking sessions.

In the winter 2008 marking session, all *CAE* examiners received face-to-face training to cover mark scheme induction as well as practical scoris® training. This was supported with ESOL-specific user guides and online training which examiners could access from home to support their initial training. As the rollout of scoris® has continued, more than 700 examiners have been trained on *Cambridge English: Preliminary (PET)*, *Cambridge English: First (FCE)*, *Cambridge English: Advanced (CAE)*, *Cambridge English: Proficiency (CPE)* and the *Business English Certificate (BEC)* suite of papers. Feedback from the training sessions was overwhelmingly positive with examiners feeling supported and prepared for live marking on the new system addressing issues raised by examiners in Raikes et al (2004) in relation to training and support being provided.

Feedback from the marking study also suggested that there were elements of the onscreen marking process which were problematic for examiners (detailed in Geranpayeh 2011 in this issue) and did not aid their marking, e.g. the use of electronic annotations on scripts. We were able to investigate these during early live marking sessions and subsequently to remove them from the process with no impact on marking quality. Similarly, the recording of marks was problematic for components with optional questions and RM was able to make technical changes to scoris® to simplify this process.

Conclusion and future directions

The flexibility of online marking allows consideration of alternative models of marking which would not be feasible for paper-based marking. scoris® has the capability to support item level marking with different sections of a Writing paper being sent to different examiners, which could improve reliability of the overall marking of the Writing component. Further functionality within the system allows for the identification of gold standard scripts to be used to 'seed' into the marking of each examiner. This involves a group of Principal Examiners or Senior Team Leaders carrying out some preliminary marking with analysis carried out on the marks awarded to identify responses where there is a very high level of agreement in marks awarded. These scripts can then be fed into the marking of all examiners. Examiners blind mark these scripts as they are not identified differently to markers. Reports on the marking of these seeding scripts can be run to show which examiners are marking within any level of tolerance set, which examiners require further support to bring their marking in line and which examiners are not able to apply the mark scheme to the required level of accuracy and should not continue marking. This process could be used to either supplement, or replace, monitoring by Team Leaders and would maintain a quality focus throughout the marking period. This is currently being considered for use as part of further improvements to the Quality Assurance process.

Online marking can be seen to be removing constraints for the continual growth which Cambridge ESOL is experiencing. In 2008 there were 251 exam sessions; while in 2012 there will be almost 700 sessions. Such growth demands increased flexibility in marking and the processing of results. As with all innovations, communication and consultation with stakeholders is key to success. Future considerations in this field will continue to take advantage of the experience and expertise which examiners hold in order to build on the success which scoris® marking has provided to date.

References

Geranpayeh, A (2011) The impact of online marking on examiners' behaviour, *Research Notes* 43 15–21.

Raikes, N, Greatorex, J and Shaw, S (2004) *From paper to screen: some issues on the way*, paper presented at International Association of Educational Assessment, June 2004, Philadelphia, USA.

RM plc: retrieved from http://www.rm.com/investors/NewsDetail.asp?cref=IN923703

The impact of online marking on examiners' behaviour

ARDESHIR GERANPAYEH RESEARCH AND VALIDATION GROUP, CAMBRIDGE ESOL

Introduction

The *Cambridge English: Advanced (CAE)* Writing paper moved to an onscreen marking system in December 2008. The present research was set up to study a number of marking issues in the new format. Issues such as usability of the onscreen marking system, comparability of paper-based marking with onscreen marking, examiners' behaviour in the two modes, and the use of the new mark scheme were of particular interest for the purposes of examiner training for the first administration of the

revised/updated *CAE* in December 2008. To achieve the above, responses of live June 2008 *CAE* candidates to a prototype *CAE* Writing paper were marked by eight examiners in both modes. The results show that the mode of marking (paper-based versus onscreen) had no impact on examiners' marking. Examiners' feedback also revealed that they had little problem in using the onscreen marking system once they received the initial training.

Background

The *CAE* Writing paper was revised in 2008. The changes in the paper included reducing the input and output in the article, report or proposal task, adding set text questions as an option in Part 2 and reducing the time for the paper by 30 minutes. For detailed descriptions of the changes see Hawkey (2009). In addition to the above changes to the format of the test, the marking of the paper was moved to an online marking system, called scoris®, that Cambridge Assessment was developing (see Cooze 2011 in this issue). Since we had no experience in using an online marking system, it was decided to set up a research project to look at a number of issues which might have an impact on examiners' marking behaviour prior to the live administration of the new test in December 2008.

Method

Research questions

There were five main lines of enquiry in the present study which investigates the impact of online marking on examiners' behaviour:

1) Were the marks given to candidates in the two marking modes comparable?
2) Did the mode of the marking result in differential severity among examiners?
3) Do the measures of consistency in the two modes of marking differ significantly?
4) What impact did the different modes have on the reliability of the examiners' marking?
5) Were the examiners using the full length of the scale in the two different modes of marking?

Data collection

A prototype *CAE* Writing paper used in the current study was constructed in the summer of 2008 and consists of two obligatory tasks. Around 200 *CAE* candidates who were taking the *CAE* June 2008 session were asked to sit this additional paper after their examination. These additional scripts, i.e. responses to the new *CAE* Writing paper, were scanned into scoris® and made available for marking. Copies of the original scripts were also saved for paper-based marking. The scripts were randomly allocated into two groups of 100 scripts each.

Eight senior examiners were invited to mark the scripts. They were also divided into two groups in scoris®. Group A Examiners (A–D) marked scripts 1–100 onscreen and then marked scripts 101–200 on paper. Group B Examiners (E–H) marked scripts 101–200 onscreen, after which they marked scripts 1–100 on paper. The order of marking mode was not controlled in this study as Johnson & Nádas (2009), who had already investigated this issue with regard to scoris®, found that it had no impact. Onscreen marking was carried out onsite in Cambridge ESOL premises, while paper marking was carried out subsequently at home, due to practical and time considerations. In normal practice, scoris® marking would also be done at home. These different marking environments were not presumed to have an impact on marking performance.

The scripts were set up in scoris® prior to marking taking place and the paper-based scripts were allocated to examiners. Each examiner marked every script, marking half of the scripts onscreen and half on paper. They did not mark the same scripts twice. Each examiner in the two groups (A–D or E–H) marked the same scripts using the same mode as the other examiners in their group. The marks awarded onscreen were captured electronically. Scripts for paper marking were copied and provided to the appropriate examiners along with suitable mark sheets to record marks before they are returned and captured electronically along with the onscreen marks.

The mark scheme used in this exercise was new, modelled on *Cambridge English: First (FCE)* and *Cambridge English: Proficiency (CPE)*. It consists of a 5-point holistic scale with three levels within each point. These are weighted to 20 marks. The examiners were also asked to annotate the scripts – yet another novelty for *CAE* examiners, but a normal practice for *FCE* and *CPE*.

Although each examiner was supposed to mark 100 scripts on paper and 100 on scoris® according to the original plan, each examiner marked 100 scripts on paper and only between 34–67 scripts on scoris®.

Results and discussion

In this section we will report the results from FACETS analysis and feedback we received from the examiners.

The total of 172 scripts was marked in scoris® by at least one examiner in the dataset. Out of this, 48 scripts were marked by all four examiners in each group. The total number of scripts marked in scoris® by each examiner, the average scores awarded and the standard deviation (SD) of those scores are presented in Table 1.

Table 2 presents the number of scripts that were marked in scoris® by all four examiners in each group.

Upon the completion of marking, a feedback questionnaire was provided to examiners and additional interviews were conducted to determine examiners' thoughts and opinions on marking in two different modes.

The data was analysed using FACETS Version 3.64 (Linacre 2006), a multi-faceted Rasch analysis program that provides estimates of examiner (or: rater) harshness and task difficulty, as well as probabilistic estimates of examinee ability and scale difficulty. FACETS reports these on the same linear scale measured in logits.

Each research question and the associated results and findings are discussed below.

Table 1: Number of scripts marked and average scores on scoris®

	Group A Examiners				Group B Examiners			
Examiner	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Total number of scripts marked	37	34	41	67	46	65	62	48
Average mark*	27.35	27.65	26.78	25.00	26.80	26.60	28.27	27.10
SD	5.42	4.95	5.60	5.17	4.34	3.95	5.75	3.38

*The maximum weighted mark is 40

Table 2: Number of scripts marked by all four examiners in each group on scoris®

	Group A Examiners				Group B Examiners			
Examiner	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Number of scripts marked by all four examiners in each group	13	13	13	13	35	35	35	35
Average mark	28.83	27.56	28.30	25.35	25.81	25.76	26.83	26.56
SD	3.02	3.24	2.80	3.28	5.38	5.07	6.25	4.86

1. Were the marks given to candidates in the two marking modes comparable?

Table 3 shows no difference between the two modes of marking: paper-based and onscreen marking in scoris® have an identical observed average of 13.4, the fair average values also being very similar (see Table 3).

Table 3: Values for mode of marking (paper-based vs. scoris®)

N	Mode	Total count	Obsvd average	Fair-M average	Measure	Infit MnSq	Outfit MnSq
1	**scoris®**	800	13.4	13.35	-0.01	0.97	0.96
2	**paper**	1,590	13.4	13.31	0.01	1.02	1.00

It can, therefore, be concluded that marks given in the two scenarios (paper and onscreen/scoris®) are very similar.

2. Did the mode of the marking result in differential severity among examiners?

FACETS provides estimates of examiner severity and consistency.

The table in Appendix 1 displays a number as an indication of the magnitude of the differences among elements of a facet, in this case, the severity among eight examiners. The Separation Index is the ratio of the corrected standard deviation (Adjusted True SD) of examiners to the root mean-square standard error (RMSE). If the examiners were equally severe, the standard deviation of the examiner difficulty estimates should be equal to or smaller than the mean estimation error of the entire dataset. The Separation Index for the sample of eight examiners is 5.66, indicating that the variance among examiners is about six times the error of the estimates. The Reliability statistic of 0.97 provided by the FACETS analysis indicates the degree to which the analysis reliably distinguishes between different levels of severity among the examiners. The Fixed chi-square tests the null hypothesis that all the elements of the facet of examiners are equal. The chi-square value of 240.5 with 7 degree of freedom (*df*) is significant at $p = .00$, indicating that the null hypothesis must be rejected, in other words, the examiners are not equally severe. All of these statistics show that eight examiners are not equal in their severity.

To examine the interaction between examiners and modes, FACETS can be used to investigate a bias interaction between these two facets. For these purposes, the facets of examiners and mode need to be constrained. Appendix 2 shows that only in the case of Examiners C and G is there a significant bias size of .19 ($p = .0105$) and -.13 ($p=.0338$), respectively, when those examiners were marking in scoris®. In all other cases, the bias size is not significant. It is important to note, however, that the size of these bias terms is negligible. For example, in the case of Examiner C, the difference between the observed and the expected score is 0.43, which is less than a sixth of a scale point.

The results of the bias interaction analyses between examiners and modes indicate that the mode of marking does not result in a significant impact on behaviour of the eight examiners. Only two examiners (mentioned above) are an exception when marking onscreen.

As a summary, Table 4 shows a graphical overview of the results. The scale along the left represents the logit (difficulty/ability) scale, which is the same for all facets: candidates, examiners/raters, mode, and tasks. It shows that the sample of scripts marked in this exercise covers a spread of ability with the most able examinees at the top and least able at the bottom of the second column. The third column in the table shows that examiners exhibit a largely similar behaviour. Examiners E, H and G are slightly more lenient than the majority, but are still within acceptable levels (-0.77 to -0.02 logits). Tasks 1 and 2 are operating at the same difficulty or tap into the same ability. The most likely scale score for each ability level is shown in the rightmost column.

As a result of the FACETS analysis, it can be seen that three out of eight examiners showed more leniency in their

Table 4: FACETS output with two marking modes in the same analysis

```
All Facet Vertical "Rulers"

Vertical = (1*,2A,3A,4A,S) Yardstick (columns lines low high extreme)=
0,7,-3,4,End
+--------------------------------------------------------------------+
|Measr|+Examinee |-Rater   |-Mode            |-Items            |Scale |
|-----+----------+---------+-----------------+------------------+------|
|   4 +          +         +                 +                  +(20)  |
|     |          |         |                 |                  | ---  |
|     |          |         |                 |                  |      |
|     |          |         |                 |                  |      |
|     |          |         |                 |                  |      |
|     |          |         |                 |                  |      |
|     | .        |         |                 |                  |  19  |
|   3 +          +         +                 +                  +      |
|     |          |         |                 |                  |      |
|     |          |         |                 |                  | ---  |
|     | *        |         |                 |                  |      |
|     |          |         |                 |                  |      |
|     |          |         |                 |                  |  18  |
|     |          |         |                 |                  |      |
|   2 + *.       +         +                 +                  + ---  |
|     |          |         |                 |                  |  17  |
|     | *        |         |                 |                  |      |
|     | *.       |         |                 |                  | ---  |
|     | ***      |         |                 |                  |  16  |
|     | **       |         |                 |                  | ---  |
|     | **       |         |                 |                  |      |
|   1 + ***      +         +                 +                  +  15  |
|     | *****    |         |                 |                  | ---  |
|     | *****    |         |                 |                  |  14  |
|     | *****    |         |                 |                  |      |
|     | *******. |         |                 |                  | ---  |
|     | *******. |         |                 |                  |      |
|     | ******.  |         |                 |                  |  13  |
*   0 * *****    * C       * Paper  scoris®  * Task 1   Task 2  * ---  *
|     | ****     | D       |                 |                  |      |
|     | *****    | A  B  F |                 |                  |  12  |
|     | ***.     | E  H    |                 |                  |      |
|     | ******   |         |                 |                  | ---  |
|     | *.       | G       |                 |                  |      |
|     | **.      |         |                 |                  |  11  |
|  -1 + **.      +         +                 +                  +      |
|     | ***      |         |                 |                  | ---  |
|     | ****     |         |                 |                  |      |
|     | ****     |         |                 |                  |  10  |
|     | *        |         |                 |                  | ---  |
|     |          |         |                 |                  |  9   |
|     | ***      |         |                 |                  | ---  |
|  -2 + *        +         +                 +                  +  8   |
|     | .        |         |                 |                  | ---  |
|     | *.       |         |                 |                  |  4   |
|     |          |         |                 |                  | ---  |
|     |          |         |                 |                  |      |
|     |          |         |                 |                  |  1   |
|     |          |         |                 |                  |      |
|  -3 +          +         +                 +                  + (0)  |
|-----+----------+---------+-----------------+------------------+------|
|Measr| * = 2    |-Rater   |-Mode            |-Items            |Scale |
+--------------------------------------------------------------------+
```

marking than their peers, but their leniency was still within acceptable limits.

3. Do the measures of consistency in the two modes of marking differ significantly?

The FACETS output in Appendix 1 also provides two measures of fit or consistency: the *infit* and the *outfit* values. The infit is the weighted mean-squared residual which is sensitive to unexpected responses near the point where decisions are being made, while the outfit is the unweighted mean-squared residual and is sensitive to extreme scores. There are no hard-and-fast rules for what degree of fit is acceptable, but lower and upper bound limits of 0.4 and 1.5 respectively for mean squares are useful and acceptable for practical purposes (Linacre 2002, Wright, Linacre, Gustafsson & Martin-Loff 1994). Fit statistics of 1.6 or greater indicate too much unpredictability in examiners' scores, while fit statistics of less than 0.4 indicate overfit or not enough variation. The infit and outfit values for eight examiners in the *CAE* study are presented in Table 5.

Table 5: Measures of variation in examiners' scores

Examiner	Total count	Obsvd average	Fair-M average	Measure	Infit MnSq	Outfit MnSq
A	274	13.3	13.10	-0.24	0.98	0.98
B	268	13.4	13.19	-0.28	1.01	0.98
C	282	12.8	12.66	-0.02	0.97	0.96
D	334	12.9	13.01	-0.20	0.99	0.99
E	292	13.5	13.48	-0.42	0.90	0.89
F	324	13.3	13.32	-0.35	0.75	0.77
G	324	14.4	14.27	-0.77	1.23	1.19
H	292	13.6	13.64	-0.50	1.17	1.16

Applying the acceptable limits to infit and outfit values between 0.4–1.5, indicating either not enough variation or too much unpredictability in examiners' scores, we can see in Table 5 that none of the examiners had very high *infit* or very low *outfit* statistics, indicating that they all showed acceptable amounts of variation but not too much unpredictability in their scores, as all of them fall within the limits of acceptable fit.

4. What impact did the different modes have on the reliability of the examiners' marking?

There is generally no single agreed index of inter-rater reliability (IRR). It depends on the purpose for which the ratings are being collected, and the philosophy underlying the rating process. Typical indexes for IRR include: proportion of exact agreements (Cohen's Kappa), correlations and variances (G-Theory). Besides, there is no clear definition of *agreement* in the previous literature, so one must decide what the term means for the testing situation or purpose of investigation.

Agreement can mean: 'To what extent do pairs of raters agree on the same rating?' This is the *exact observed agreement* statistic. If we wish our raters to act like 'rating machines', we expect to see agreement of 90%+. Raters are often trained to act like this. In addition, agreement can mean: 'Are the ratings of pairs of raters highly correlated?' FACETS does not report this directly, but correlations between raters can be calculated. Agreement can also be taken to mean: 'Are pairs of raters acting like independent experts?' If they are, the observed agreements will be close to the expected agreements.

One can also ask 'Do raters have the same level of leniency/severity?' This is reported in FACETS as the *Reliability (not inter-rater)* statistic. In FACETS, this statistic needs to be close to 0, so that the rater measures are not reliably different. Also the *Fixed all-same* chi-square test is expected to not be rejected so that we have evidence that raters are not behaving differently.

For the present study, the relevant meaning of rater agreement is to treat raters as independent experts, possibly showing variation in conditions thought to be identical. FACETS models raters to be 'independent experts'. In FACETS, an inter-rater reliability coefficient, IRR, is not computed, however, from one perspective, it is the reverse of the Separation Reliability, i.e. '1 – Separation Reliability'.

If raters must agree on the exact value of the ratings, Cohen's-Kappa type of inter-rater reliability index needs to be used.

Cohen's Kappa is *(Observed Agreement % – Chance Agreement %) / (100–Chance agreement %)*, where chance is determined by the marginal category frequencies.

A Rasch version of this would use the *Expected Agreement %* for an adjustment based on 'chance + rater leniency + rating scale structure'. Then the Rasch-Cohen's Kappa would be:

Equation 1

$$\text{Rasch-Cohen's Kappa} = \frac{(\text{Observed agreement \%} - \text{Expected agreement \%})}{(100 - \text{Expected agreement \%})}$$

Under Rasch-model conditions, this should ideally be close to 0.

To see whether Rasch-Cohen's Kappa is close to 0 in the *CAE* study, the following information in Table 6 was used from one of the tables of the FACETS Output analysis.

Table 6: Exact and expected examiner agreements

Inter-examiner agreement opportunities	3,176	
Exact agreements	696	21.9%
Expected agreements	608.1	19.1%

Table 7a: Correlations between examiners' marking in the paper-based mode

	Group A correlations			
	A	**B**	**C**	**D**
A				
B	0.77			
C	0.78	0.76		
D	0.78	0.80	0.76	
	Group B correlations			
	E	**F**	**G**	**H**
E				
F	0.69			
G	0.82	0.79		
H	0.68	0.77	0.72	

Table 7b: Correlations between examiners' using scoris® marking

	Group A correlations			
	A	**B**	**C**	**D**
A				
B	0.82			
C	0.88	0.86		
D	0.83	0.77	0.82	
	Group B correlations			
	E	**F**	**G**	**H**
E				
F	0.83			
G	0.65	0.80		
H	0.67	0.68	0.67	

According to the figures in Table 6, the value of Rasch-Cohen's Kappa for the *CAE* dataset is:

$$\text{Rasch-Cohen's Kappa} = \frac{21.9-19.1}{100-19.1} = \frac{2.8}{80.9} = 0.0346$$

Thus, the Rasch-Cohen's Kappa for the reliability of the marking of eight independent examiners in the *CAE* data using Equation 1 is 0.0346. This is close enough to zero, so inter-rater reliability is within the acceptable range in this study.

Tables 7a and 7b show Pearson correlations between eight examiners in the two groups and two modes (paper-based and onscreen in scoris®).

It can be seen that correlations are not very high within the second group of examiners: between Examiners E and F in paper-based marking (see Table 7a) and Examiners E, F, G and H in scoris® (see Table 7b). These low correlations, i.e. in the high 0.60s, are confirmed by the FACETS analysis, graphically presented in Table 4, showing that these examiners were found to be the most lenient, relative to the others in the examiners group. This indicates that the agreement between a quarter of the examiners could be improved.

5. Were the examiners using the full length of the scale in the two different modes of marking?

When examining examiner behaviour in the two modes separately, it was found that it does not change significantly in terms of severity/leniency relative to each other. The only difference in examiner behaviour, shown in the last two columns in Table 8, is the slight difference in the most likely scale score awarded for each ability level in the two marking modes (1=scoris®, 2=paper). When marking on paper, examiners awarded scores on a slightly shorter scale, especially at the extremes of the ability scale. However, in the middle of the scale the bands for the most likely scores overlap exactly in the two modes, i.e. scores of 9, 10, 11 and 14, 15 were awarded to candidates of the same ability range in both modes. So, there is evidence that in scoris® markers use the extremes of the full scale more readily than when marking on paper.

Examiners' feedback

In the feedback questionnaire, four examiners reported feeling apprehensive or worried about the new marking mode onscreen, but three said they kept an open mind and one reported looking forward to it. However, marking in long periods onscreen caused problems for some of them. For instance, the use of the mouse and other physical strain led to a feeling of multitasking. Examiners' confidence in their own marking accuracy was lower in scoris® due to the constant need to scroll up and down the screen to recall information. They reported that completing the marks on scoris® was frustrating. In general, five of them said that marking on scoris® was slow, with two claiming it was very slow and one suggesting it may be faster. However, examiners were comparing their marking speed to the speed at which they normally marked paper-based *CAE*.

Examiners' confidence in their own marking consistency was generally higher when they began to work with scoris®

Table 8: FACETS output with two modes (paper-based and onscreen) presented separately

```
All Facet Vertical "Rulers".

Vertical = (1*,2A,3A,4A) Yardstick (columns,lines,low,high)= 0,7,-3,4
--------------------------------------------------------------------------------------
|Measr |+Examinee |-Rater    |-Mode                |-Items               | S.1 | S.2 |
--------------------------------------------------------------------------------------
+    4 +          +          +                     +                     +(20) +(20) +
|      |          |          |                     |                     | --- | --- |
|      |          |          |                     |                     |     |     |
|      |          |          |                     |                     |     |     |
|      |          |          |                     |                     |     |     |
|      |          |          |                     |                     |     |     |
|      | .        |          |                     |                     |  19 |     |
+    3 +          +          +                     +                     +     +  19 +
|      |          |          |                     |                     |     |     |
|      |          |          |                     |                     | --- |     |
|      | *        |          |                     |                     |     | --- |
|      |          |          |                     |                     |  18 |     |
|      |          |          |                     |                     |     |  18 |
|      |          |          |                     |                     | --- |     |
+    2 + *        +          +                     +                     +  17 + --- +
|      | .        |          |                     |                     |     |  17 |
|      | *        |          |                     |                     | --- |     |
|      | .        |          |                     |                     |  16 | --- |
|      | ***.     |          |                     |                     |     |  16 |
|      | **.      |          |                     |                     | --- | --- |
|      | *.       |          |                     |                     |  15 |     |
+    1 + ***.     +          +                     +                     +     +  15 +
|      | *****    |          |                     |                     | --- | --- |
|      | *****    |          |                     |                     |  14 |  14 |
|      | *****.   |          |                     |                     |     |     |
|      | *******. |          |                     |                     | --- | --- |
|      | *******. |          |                     |                     |     |  13 |
|      | ******   |          |                     |                     |  13 |     |
*    0 * ****.    * C        * Paper   scoris®     * Task 1   Task 2     *     * --- *
|      | *****.   | D        |                     |                     | --- |     |
|      | ****.    | A   B   F|                     |                     |     |  12 |
|      | ***.     | E        |                     |                     |  12 |     |
|      | *****    | H        |                     |                     | --- | --- |
|      | **.      | G        |                     |                     |     |     |
|      | *.       |          |                     |                     |  11 |  11 |
+   -1 + ***      +          +                     +                     +     +     +
|      | ***      |          |                     |                     | --- | --- |
|      | *****    |          |                     |                     |  10 |     |
|      | **.      |          |                     |                     |     |  10 |
|      | *.       |          |                     |                     | --- | --- |
|      |          |          |                     |                     |  9  |  9  |
|      | **.      |          |                     |                     | --- | --- |
+   -2 + *.       +          +                     +                     +     +  8  +
|      | .        |          |                     |                     |  8  | --- |
|      | *.       |          |                     |                     | --- |  4  |
|      |          |          |                     |                     |  4  | --- |
|      |          |          |                     |                     | --- |  1  |
|      |          |          |                     |                     |     |     |
|      |          |          |                     |                     |  1  | --- |
+   -3 +          +          +                     +                     + (0) + (0) +
--------------------------------------------------------------------------------------
|Measr | * = 2    |-Rater    |-Mode                |-Items               | S.1 | S.2 |
--------------------------------------------------------------------------------------
```

more often. However, they did report that they found it easier to navigate in scoris® than they had anticipated previously. They found it very easy to read onscreen, perhaps even easier than while doing paper and pencil marking and considered the possibility of zooming helpful. They would find it useful to have shortcut keys and a drawing tool constantly present on the screen.

These results of the qualitative data collection (feedback questionnaire and interview) largely corroborate the findings of other research studies on onscreen marking within Cambridge Assessment (e.g. Johnson & Nádas 2009). After an initial resistance to the new marking mode, examiners found the software easy to use and their confidence in their marking accuracy grew with practice. However, there are physical strains to be considered (caused by long periods of marking onscreen). Annotations were not seen as a very useful aspect of the software. The findings from this study have helped to improve the final version of scoris®.

Conclusions and future research

The aim of this study was to investigate the impact of using an online marking system on examiners' behaviour. The results show that the examiners had little problem in using the onscreen marking once they received the initial training. They also show that the mode of marking (paper-based vs. onscreen) had no impact on examiners' marking.

We started out with the following research questions:

1) Were the marks given to candidates in the two marking modes comparable?
2) Did the mode of the marking result in differential severity among examiners?
3) Do the measures of consistency in the two modes of marking differ significantly?
4) What impact did the different modes have on the reliability of the examiners' marking?
5) Were the examiners using the full length of the scale in the two different modes of marking?

The analysis carried out on the *CAE* Writing scripts showed that the marks are very similar within the two scenarios (marking on paper and marking onscreen using scoris®) and across the two tasks (Q1). Even though there is a difference in severity, the rating behaviour of eight examiners is similar and entirely consistent across the two modes. The format of the marking does not result in a significant impact on examiners' behaviour, one examiner being an exception when marking onscreen (Q2). No examiner misfitted in either mode, which means their rating was within acceptable/tolerable limits in terms of consistency. In other words, none of the examiners had very high infit or very low outfit statistics, indicating that they all showed acceptable amounts of variation, but not too much unpredictability in their scores (Q3). The inter-rater reliability of the examiners' rating is acceptable, even though the agreement within a quarter of the pairs of examiners could be improved (Q4). With a slight difference, the examiners seem to be using the same length of the scale in the two modes of marking, especially in the middle of the scale. There is evidence that at the extreme ends of the scale they are more likely to stretch the bands and use the full range of marks when marking onscreen (Q5).

Looking ahead, it would be useful to explore whether markers attend to the same features of writing performance in paper-based and computer-based marking. This data could be gathered through successive trials of onscreen marking within Cambridge ESOL as well as with the help of other related software developments. This strand of research is a candidate for further investigation within the English Profile research on the Cambridge Learner Corpus where salient linguistic features of learner performance are identified at each CEFR level. One of the applications could be a list of useful features that could be made available to examiners in the software or in the mark schemes in an attempt to harmonise the features examiners attend to while marking in the two marking modes (paper and onscreen).

References

Cooze, M (2011) Assessing Writing tests on scoris®: The introduction of online marking, *Research Notes* 43, 12–15.

Hawkey, R (2009) *Examining FCE and CAE*, Studies in Language Testing series, volume 28, Cambridge: UCLES/Cambridge University Press.

Johnson, M and Nádas, R (2009) Marking essays onscreen: an investigation into the reliability of marking extended subjective texts, *British Journal of Educational Technology* 41 (5), 814–826.

Linacre, J M (2002) What do infit, outfit, mean-square and standardized mean? *Rasch Measurement Transactions* 16, 878.

Linacre, J M (2006) *Facets Rasch measurement computer program*, Chicago: Winsteps.com.

Wright, B D, Linacre, M, Gustafsson, J E, and Martin-Loff, P (1994) Reasonable mean-square fit values, *Rasch Measurement Transactions* 8 (3), 370.

Appendices

Appendix 1: Reliability of rater measurement

CAE Equivalence trial Study January 21, 2009. 01-21-2009 15:35:29
Table 7.2.1 Rater Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair-M Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Obs %	Agree. Exp %	N	Rater
3649	274	13.3	13.10	-.24	.04	.98	-.1	.98	-.1	1.00	.84	.09	20.7	19.5	1	A
3594	268	13.4	13.19	-.28	.04	1.01	.1	.98	-.1	1.03	.84	.09	20.2	19.3	2	B
3606	282	12.8	12.66	-.02	.04	.97	-.3	.96	-.4	.99	.84	.10	23.9	19.2	3	C
4322	334	12.9	13.01	-.20	.04	.99	-.1	.99	.0	.98	.82	.10	20.8	19.5	4	D
3941	292	13.5	13.48	-.42	.04	.90	-1.2	.89	-1.3	1.12	.82	.09	22.7	19.2	5	E
4311	324	13.3	13.32	-.35	.04	.75	-3.4	.77	-3.0	1.20	.84	.10	22.0	19.0	6	F
4666	324	14.4	14.27	-.77	.04	1.23	2.7	1.19	2.3	.84	.83	.10	22.3	18.3	7	G
3985	292	13.6	13.64	-.50	.04	1.17	1.9	1.16	1.7	.84	.75	.10	22.6	19.3	8	H
4009.3	298.8	13.4	13.33	-.35	.04	1.00	-.1	.99	-.2		.82				Mean (Count: 8)	
369.2	23.6	.5	.45	.21	.00	.14	1.8	.13	1.6		.03				S.D. (Population)	
394.7	25.2	.5	.48	.22	.00	.15	1.9	.14	1.7		.03				S.D. (Sample)	

Model, Populn: RMSE .04 Adj (True) S.D. .21 Separation 5.29 Reliability (not inter-rater) .97
Model, Sample: RMSE .04 Adj (True) S.D. .22 Separation 5.66 Reliability (not inter-rater) .97
Model, Fixed (all same) chi-square: 240.5 d.f.: 7 significance (probability): .00
Model, Random (normal) chi-square: 6.8 d.f.: 6 significance (probability): .34
Inter-Rater agreement opportunities: 3176 Exact agreements: 696 = 21.9% Expected: 608.1 = 19.1%

Appendix 2: Interaction between rater and mode

CAE Equivalence trial Study January 21, 2009. 01-21-2009 15:35:29
Table 13.1.1 Bias/Interaction Calibration Report (arranged by N).

Bias/Interaction: 2. Rater, 3. Mode (higher score = higher bias measure)

Obsvd Score	Exp. Score	Obsvd Count	Obs-Exp Average	Bias Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	Rater Sq	Rater N	Rater R	Rater measr	Mode N	Mode	Mode measr
1012	1007.8	74	.06	.03	.08	.33	73	.7448	.7	.7	1	1	A	-.24	1	scoris	.00
940	941.2	68	-.02	-.01	.08	-.10	67	.9228	1.0	1.0	2	2	B	-.28	1	scoris	.00
1098	1062.5	82	.43	.19	.07	2.62	81	.0105	.8	.8	3	3	C	-.02	1	scoris	.00
1675	1704.9	134	-.22	-.10	.06	-1.72	133	.0874	1.0	1.0	4	4	D	-.20	1	scoris	.00
1233	1221.5	92	.13	.06	.07	.81	91	.4194	.7	.7	5	5	E	-.42	1	scoris	.00
1729	1709.1	130	.15	.07	.06	1.18	129	.2391	.7	.7	6	6	F	-.35	1	scoris	.00
1753	1788.7	124	-.29	-.13	.06	-2.15	123	.0338	1.3	1.3	7	7	G	-.77	1	scoris	.00
1301	1288.0	96	.14	.06	.07	.90	95	.3696	1.3	1.3	8	8	H	-.50	1	scoris	.00
2637	2641.1	200	-.02	-.01	.05	-.20	199	.8452	1.1	1.1	9	1	A	-.24	2	paper	.00
2654	2652.7	200	.01	.00	.05	.06	199	.9514	1.0	1.0	10	2	B	-.28	2	paper	.00
2508	2543.5	200	-.18	-.08	.05	-1.69	199	.0928	1.0	1.0	11	3	C	-.02	2	paper	.00
2647	2617.0	200	.15	.07	.05	1.43	199	.1545	.9	.9	12	4	D	-.20	2	paper	.00
2708	2719.4	200	-.06	-.03	.05	-.54	199	.5903	1.0	1.0	13	5	E	-.42	2	paper	.00
2582	2601.8	194	-.10	-.05	.05	-.95	193	.3424	.8	.8	14	6	F	-.35	2	paper	.00
2913	2877.1	200	.18	.08	.05	1.70	199	.0912	1.1	1.1	15	7	G	-.77	2	paper	.00
2684	2696.9	196	-.07	-.03	.05	-.62	195	.5378	1.1	1.1	16	8	H	-.50	2	paper	.00
2004.6	2004.6	149.4	.02	.01	.06	.07			1.0	1.0	Mean (Count: 16)						
701.9	705.4	52.3	.17	.08	.01	1.29			.2	.2	S.D. (Population)						
724.9	728.5	54.0	.18	.08	.01	1.34			.2	.2	S.D. (Sample)						

Fixed (all = 0) chi-square: 26.8 d.f.: 16 significance (probability): .04

The BULATS Online Speaking Test

LUCY CHAMBERS RESEARCH AND VALIDATION GROUP, CAMBRIDGE ESOL
KATE INGHAM ASSESSMENT AND OPERATIONS GROUP, CAMBRIDGE ESOL

Introduction

Computer-based (CB) language assessment has grown considerably in importance in the last few decades. However, although Cambridge ESOL produced its first CB-based test in the mid 1990s, the use of technology in the assessment of Speaking is a relatively new area of CBT for the organisation. This article outlines aspects of the development of the *BULATS Online Speaking Test*, with a specific focus on a proof-of-concept trial and alignment to the Common European Framework of Reference (CEFR) (Council of Europe 2001). These activities have formed part of the validation of this new test.

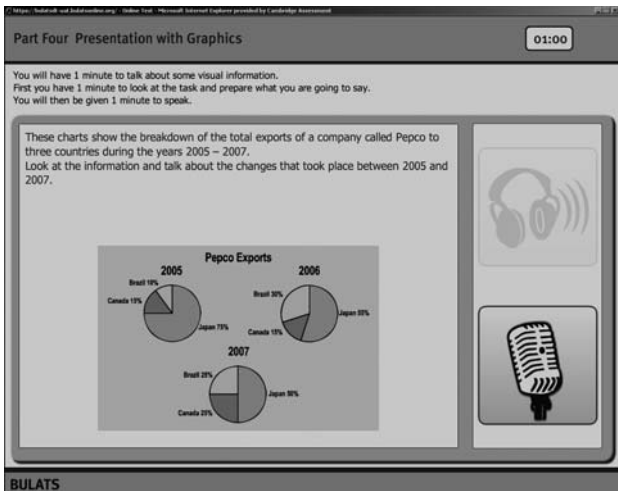
BULATS (Business Language Testing Service) is an assessment tool of language in a work context developed by Cambridge ESOL in response to the need for a reliable, efficient, flexible and easy to administer test for company use. *BULATS* is a multilingual test and assesses from A1 to C2 on the CEFR. A *BULATS Online Reading and Listening Test*

has been available for some time and there has been a growth in demand for online versions of the Speaking and Writing tests to complement this provision and allow online testing of all four language skills.

Format of the *BULATS Online Speaking Test*

The *BULATS Online Speaking Test* has five parts. In Part 1, test takers respond to eight questions about themselves and their work (e.g. How do you use English in your job?). Part 2 involves the repetition of text that might be read aloud in a work or business situation. In Part 3, the candidate talks about a work-related topic (e.g. The perfect office) with the help of prompts which appear on the screen. Part 4 (illustrated in Visual 1) involves delivering a mini-presentation describing a visual such as pie charts or a bar chart related to a business situation (e.g. Company exports). Part 5 requires test takers to imagine they are in a specific situation with another person (e.g. a colleague) and have to respond to questions that may be asked in that situation (e.g. advice about planning a conference).

Visual 1: Screenshot of Part 4 of the BULATS Online Speaking Test



Before starting the test, candidates view an online tutorial which outlines the format of the test and provides examples of the tasks that will be included. The candidate wears a headset with a microphone attached. Questions are then presented to or heard by the candidate via the computer and the candidate's responses are recorded. These recordings are then accessed and assessed by examiners after the test.

The *BULATS* online application was also developed for training, standardisation and certification of examiners before they become eligible to rate Speaking test performances. This ensures that all examiners, wherever they are in the world, undergo the same activities and meet the same standards. Once the examiner starts rating, their performance is monitored via the online platform and further training and support is made available as required.

Trial of the Online Speaking Test

An initial small-scale feasibility trial was set up in four countries (UK, India, Argentina and Switzerland) to investigate technical issues and to gather test takers' perceptions of the test. The trial had five specific objectives:

- to gather candidate feedback on the screen designs
- to collect feedback on test content, format and length in order to inform future development of the test
- to evaluate whether the Online Speaking Test could be delivered to a number of candidates simultaneously, in a secure environment without any compromise on the quality of audio (not discussed here)
- to test whether candidate speech captured without an examiner in real time can be reliably assessed
- to determine whether candidates' responses could reliably be recorded and stored in real time and accessed remotely.

The initial candidate trial involved 100 participants (53 females and 47 males) with ages ranging from 19 to 53, with 87% falling into the 25–40 age bracket. The sample came from 21 different countries and included 19 different first languages; 90% of candidates were in work or had work experience; the remaining 10% were enrolled on Business English courses. On a self-rating scale of language proficiency, they ranged from A1 to C2 on the CEFR (2001).

Extensive trialling of the application with examiners took place in the UK, although the examiner application itself was trialled more widely. Six examiners (two male and four female) participated in the UK trial, with ages ranging from 35 to 65 and a minimum length of examining experience of 10 years.

Candidates and examiners participating in the trial were asked to complete a comprehensive questionnaire which elicited their feedback on a number of key areas. Those taking part in the UK also participated in informal focus groups where their views on some key issues were explored.

Candidate feedback

The online application

An important consideration was the extent to which the user-interface design of the test and particularly the use of the timer (which showed test takers how long they had to prepare for tasks or to answer questions) would be found helpful by the candidates. The primary aim of good interface design is to minimise construct-irrelevant variance that could be attributed to test method (Messick 1989). Fulcher (2003) has described the importance of a number of key criteria relating to interface design such as clear text features (e.g. 12 point font size and avoidance of upper case), fast and easy navigation, simple and consistent terminology and helpful icons.

At the start of the test, the candidate is asked to carry out a voice recording to check that their voice has been recorded correctly. The incidence of 'popping' (which may be described as a 'hissing' sound produced when the candidate's mouth is too close to the microphone), however, resulted in the development of an enhanced voice check test in the next stage of the development which allows the candidate to record, listen to and re-record their voice until satisfied with the recording quality.

The majority of respondents (90%) agreed or strongly agreed that they knew when to start and stop speaking, supported by comments such as 'It's very easy to know when to speak thanks to the icons representing the microphones and headset'. They also reported that the instructions were very clear. As far as the timer was concerned, however, there was a greater division of opinion with respondents recognising the usefulness of the timer, but also its capacity to distract. The representative comments were the following: 'The timer is helpful, but in some cases I prepared a shorter answer than required' and 'I found it was kind of hard to pay attention to the clock when answering questions'.

One issue in CBT design relates to the role of computer familiarity and its possible impact on candidate performance (Taylor, Jamieson, Eignor & Kirsch 1998). Lack of computer familiarity may introduce construct-irrelevant variance in CB tests and influence scores due to the key role computer familiarity plays in test performance. In the *BULATS Online Speaking Test*, the candidate has no interaction with the keyboard other than to enter personal details at the start of the test, thus potentially reducing the effect of computer familiarity. In addition a demo test available on the website and a tutorial that can be taken prior to the test, ensure that candidates are familiar with both the test format and how they should progress through the test.

Content, topic and test length

A clear majority of respondents (71%) considered that the tasks in the test were similar to those they could be asked to do at work. The *BULATS Online Speaking* application allows for the production of specific tests for specific test populations or clients, and it would be possible for pre-service versions of the test or versions with specific domain content to be created. As the online application also captures details of candidate background, including work experience, via the Candidate Registration screen which each candidate must complete at the start of the test, it will be possible to monitor performance on the test against candidate characteristics to confirm that candidates without work experience are not disadvantaged.

As far as test length is concerned, respondents agreed, in general, that time allowed was appropriate for Parts 1 and 2 but not long enough for Parts 3, 4 and 5. This may be accounted for, however, by the fact that respondents had not been prepared for the test. An argument in favour of not increasing the duration of Parts 3, 4 and 5 was accessibility to lower level candidates – a long turn of 1 minute on any topic is already challenging for candidates at A1 and A2 levels on the CEFR, but is considered important to include in order to discriminate between candidates at the higher levels of proficiency.

Feedback to the questionnaire showed that 67% of respondents agreed or strongly agreed with the statement 'The test gave me full opportunity to demonstrate my English language skills'. However, 45% of the sample expressed the desire to be able to hear the questions again. The option for candidates to hear questions again was given consideration, but ruled out on assessment grounds, given that *BULATS* is a multi-band test assessing from A1 to C2 on the CEFR.

Taking a computer-delivered test

A further issue relates to candidate performance in a test without an examiner being present. One of the advantages of CB-based Speaking tests is high reliability due to the standardisation of test prompts and delivery. Each prompt is delivered in the same way, regardless of where the candidate takes the test. Most Cambridge ESOL test takers, however, have greater familiarity with face-to-face Speaking tests where the examiner delivers prompts in person. This is why investigating candidate perception in this area was considered important. The majority of respondents (60%) agreed or strongly agreed that they felt comfortable doing a Speaking test without an examiner and liked taking the test on computer. However, they were fairly evenly split as to their *preference* for taking the exam with or without an examiner. Some felt that the support of an Interlocutor would have had a positive impact on their performance, particularly because the Interlocutor script allows rephrasing of questions if test takers misunderstand them. One of the representative comments was the following: 'If it was face-to-face, it would be warm and comfortable – less stressful. It's nice to see someone giving reactions to your answer'. This in itself is not an untypical comment from candidates who may interpret a smile from an examiner as confirmation of good performance. Representative comments from respondents preferring a test without an examiner are the following: 'I can concentrate more than with an examiner. If I talk with an examiner, I get nervous' and 'The Online test is better because you don't know which teacher (i.e. examiner) you will get.' In general respondents were positive about the *BULATS Online Speaking Test*. They felt comfortable taking the test without an examiner present and could see the benefits that this kind of assessment can bring. Although some did express caution about not interacting with an examiner, 79% of respondents said they would like to take a test like this in the future.

Examiner feedback on the online application

As noted earlier, the trial aimed to gather, not just candidate, but also examiner feedback. The majority of Cambridge ESOL Oral Examiners are used to face-to-face tests, so it was important to gather feedback both on using the online application and on whether they felt they could confidently rate candidates using recorded performances.

Candidate recordings were accessed remotely after the tests had taken place. Examiners were positive about the experience of online marking; all respondents agreed or strongly agreed that they liked rating on computer. Furthermore, all examiners reported feeling comfortable rating the candidates online without face-to-face interaction. Two examiners expressed surprise that they would feel like this but qualified it by noting that the online Speaking examining experience involves assessment only rather than the combined roles of Assessor and Interlocutor as in the face-to-face *BULATS* test.

All examiners reported feeling confident in their ability to assess candidates fairly and accurately, even in test recordings where popping was evident. A number of candidate recordings proved to be too faint for examiners to

hear and assess. The issue here was found to be a problem with the microphone settings on a particular computer used by these candidates. To resolve these issues, optimum settings have been investigated and a diagnostic type tool developed so that centres can ensure that computers are set to the optimum level for performance before a candidate takes a test.

Concerning the application design and ease of navigation for the user, all examiners reported that they found the application easy to use. The application allows an examiner to see all the candidate responses they have been allocated; they can also click on a response and see the candidates' view and hear the response. There is an option to pause and play back the response. The examiners did, however, make a number of suggestions for improvements to the examiner screens to facilitate and streamline speed of the marking process. These suggestions fed into the subsequent developments which resulted in:

a) the reduction in the number of clicks of the mouse needed to navigate between parts of the test and
b) greying out of items after a part has been marked to allow examiners to track more easily (if necessary) where they are in the marking process.

Alignment to the CEFR

In addition to investigating aspects of the design of the test, tasks and related assessment issues, it was also considered important to gather validity evidence to support the process of alignment to the CEFR. This evidence is built up as the test is designed, developed, monitored and reviewed. The paper will describe some of the CEFR-related activities carried out during the design and development stages of *BULATS Online Speaking*, including a standard-setting exercise.

CEFR alignment evidence should not only focus on ensuring standard-setting exercises are conducted rigorously and appropriately, but also that standards are maintained across time. A number of activities are in place to support maintenance of test standards: the development and trialling of assessment criteria, the trialling of all test material and the training, standardisation and continuous monitoring of examiners (see earlier).

As part of the development of the new Speaking test, new assessment scales were created, covering the full range of the CEFR. These were written by an experienced consultant who has worked extensively with the CEFR. The individual descriptors within each band were also mapped to the CEFR assessment scales to ensure that the new scales accurately reflect the CEFR (see Chambers 2009). All test material for the Speaking test is trialled on a number of candidates and feedback collected both from the candidate and the person conducting the trialling. Evidence is sought on whether the tasks:

a) are at the appropriate level
b) are suitable for all candidates and
c) reflect the aims set out in the test specifications (e.g. in terms of content, suitability and cognitive skills).

All personnel involved in trialling will have undergone CEFR familiarisation activities.

Familiarisation with the CEFR, its levels and illustrative descriptors, ensures that all personnel involved in test specifications, test construction, examining and standard-setting activities have the knowledge to make informed decisions. It is a vital step before engaging in any CEFR-related activity. At Cambridge ESOL, a thorough set of both face-to-face and self-guided activities via the induction process and ongoing training helps ensure adequate familiarisation is achieved (see Khalifa & ffrench 2009). For *BULATS*, in addition to these routine activities, a series of familiarisation activities were completed by panellists prior to and as part of a standard-setting exercise (see below).

Standard setting

A standard-setting exercise for BULATS Online Speaking was conducted in August 2009 (Chambers, Khalifa, Walker & Fernand 2010). The purpose of this study was to establish evidence of the extent of speaking score alignment to the CEFR levels and to provide a range of recommended cut scores corresponding to the six levels of the CEFR using a benchmarking methodology. A modified Analytical Judgment method (Plake & Hambleton 2000) was used after consideration of the nature of the tasks and the panellists. This method provides panellists with an opportunity to review examinees' work and not simply estimate performance based on a scoring guide or rubric.

Ten participants (seven females and three males) were panellists in this study. They were selected based on their familiarity with the CEFR, Cambridge ESOL level-based examinations such as *Main Suite* and *Business English Certificates* (*BEC)* or *BULATS*, and their experience in the field of language testing. After familiarisation and CEFR standardisation exercises (using CEFR illustrative samples for spoken performance), the panellists were asked to classify *BULATS Online Speaking* candidates' performances into defined categories based on CEFR levels. Samples used were collected during piloting of the online system. Classification was initially carried out at a broad level (i.e. A1, A2, B1, B2, C1, or C2), and then refined by identifying the highest and lowest performances at each level. Cut scores/bands were then calculated by replacing the judgements with the actual scores each performance received and averaging these values across panellists.

The Online *BULATS* assessment scales contain six bands, each band corresponding to a CEFR level. Table 1 compares the cut-off bands derived in the workshop to these bands. On the whole, the recommended boundaries agree with the *BULATS* bands. The results should be seen as positive and as a confirmation of the tests' alignment to the CEFR. It is hypothesised that the discrepancy in cut-off at C2 was due to an insufficient sample size at this level. It must also be remembered that the CEFR scales are multi-purpose and do not capture some elements of the Online *BULATS* assessment scales such as task achievement and pronunciation. Similarly, the CEFR scales are not necessarily appropriate for use with all *BULATS* task types e.g. reading aloud.

Table 1: Comparison of preliminary BULATS bands and standard-setting cut-offs

BULATS bands	CEFR level	Workshop cut-offs
1.0–1.9	A1	
2.0–2.9	A2	2
3.0–3.9	B1	3
4.0–4.9	B2	4
5.0–5.8	C1	5
5.8–6.0	C2	5.5

The benchmarking exercise confirmed the use of the CEFR levels and cut scores that were intended when designing the test. The results of this study coupled with validation evidence on assessment scales (see Chambers 2009) give support to the relationship between *BULATS Online Speaking* and the CEFR. This means that the results awarded to candidates reflect the language abilities detailed. The intention is to repeat the standard-setting exercise after the test has been live for some time.

Conclusion

This paper described some of the activities conducted during the design and development phases of the *BULATS Online Speaking Test*. As with designing any new test, it was important to consult with various stakeholders in order to design a test that was both useful and fit for purpose. Going forward, it is important to continue examining validity by monitoring and evaluating the test once it goes live. Planned activities include gathering information on test-takers' and other test-users' perceptions of the new test.

References

Chambers, L (2009) Using the CEFR to inform assessment criteria development for Online BULATS speaking and writing, *Research Notes* 38, 29–31.

Chambers, L, Khalifa, H, Walker, C, and Fernand, S (2010) *Aligning Online BULATS Speaking & Writing to the CEFR: an exploratory study*, internal Cambridge ESOL report.

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

Council of Europe (2009) *A Manual for Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR)*, Strasbourg: Language Policy Division.

Fulcher, G (2003) Interface design in computer-based language testing, *Language Testing* 20 (4), 384–408.

Khalifa, H & ffrench, A (2009) Aligning Cambridge ESOL examinations to the CEFR: issues and practice, *Research Notes* 37, 10–14

Messick, S A (1989) Validity, in Linn, R L (Ed.) *Educational Measurement*, New York: American Council on Education/ MacMillan Publishing Company, 13–103.

Plake, B S, and Hambleton, R K (2000) A standard-setting method designed for complex performance assessments: Categorical assignments of student work, *Educational Assessment* 6 (3), 197–215.

Taylor, C, Jamieson, J, Eignor, D and Kirsch, I (1998) *The Relationship between Computer Familiarity and Performance on Computer-Based TOEFL Test Tasks*, TOEFL Research Report ETS.

Composition and revision in computer-based written assessment

LUCY CHAMBERS RESEARCH AND VALIDATION GROUP, CAMBRIDGE ESOL

Introduction

With the advent of computer-based (CB) assessment, innovative ways of recording and analysing data have emerged. At Cambridge ESOL, for example, snapshots/ backups of a candidate's CB writing output are taken at regular time intervals as part of the Cambridge Connect test delivery system (see Seddon 2005). These can be used in a research context to build up a picture of exactly how the writing text was developed. This can give information on time taken for planning, revising and writing, and at what stages different activities occurred. The use of snapshots has the added advantage of not affecting or altering the behaviour of the candidate.

When composing on computer it is easier to make revisions to the text, such as replacing, inserting, editing and deleting characters, than when writing on paper. These changes can also be made without impinging on the appearance of the text. In addition, composition on computer does not need to occur in a linear, start-to-finish, fashion; writers can build text up from notes or write the body of the text and then come back to the introduction/conclusion. Thus one issue that arises is whether or not the candidates composing on computer are able to optimise the advantages of the mode and thus produce more polished and perhaps better composition.

This paper will describe an exploratory study looking at composition and revision in a small sample of candidates who took Cambridge ESOL's *Business English Certificate (BEC) Vantage* in the CB mode. Using the snapshot technology described above, the author built a picture of text development during a live examination in an attempt to establish whether candidates had optimised the mode of administration in writing their assessment response. In particular, attention focused on the time spent on different composition/revision activities and explored the relationship between these and scores in the Writing test.

Results can provide information for teachers, learners and candidates on successful revision activities.

Literature review

Revision

Revision involves analysing what has been written and evaluating its success in conveying the intended message (Crawford, Lloyd & Knoth 2008:109), essentially identifying a mistake/weakness and then rectifying it. Successful revision 'results not from the number of changes a writer makes but from the degree to which revision changes bring the text closer to fitting the demands of the text' (Faigley & Witte 1981:411). Changes to a text can involve surface-level editing (e.g. spelling, punctuation, formatting, changing a word) or deeper-level changes where the message itself is changed (e.g. focusing on organisation and coherence). Much of the research literature has classed both of these changes as revision; although as Worden (2009:160) noted, this could be seen as too inclusive as it would consist of features traditionally seen as editing.

Revision can be affected by a number of factors such as writing experience, first language (L1), composing in a second language (L2), knowledge of revision strategies and when to apply them, context of composition (e.g. classroom, assessed, timed, importance, mode etc.) and familiarity with the task, topic, audience etc. First language studies have found that inexperienced writers focus on surface changes, mainly single-word, often lexical, revisions (Somers 1980:382), with few revisions resulting in a change of meaning (Faigley & Witte 1981:407), whereas expert writers tend to revise at all levels (Somers 1980:386). Studies comparing revision by administration mode have found a number of differences. In a study on non-native speakers, in which participants wrote two comparable writing tasks in non-timed conditions, Li (2006:5) found that participants revised more at both higher- (at or above the phrase level) and lower- (word or character) levels when composing CB texts and did less pre-planning. Van Waes & Schellens (2003:848) found that CB composers tended to revise more extensively at the beginning of the writing process, did not normally undertake any systematic revision of their work before finishing and focused on lower level linguistic features. When writing in a foreign language, language proficiency is likely to have an effect on revising processes, which may be reflected in the kinds of revisions writers make to their foreign language texts (Stevenson, Schoonen & de Glopper 2006:202). Plakans (2008:113) argues that writing expertise in L1 has a strong impact on the L2 writing process and that this impact is separate from second language (L2) proficiency and perhaps more central.

For studies examining the composition and revision processes, results have been inconsistent; this is due to different research contexts: L1 or L2 based, timed or untimed, assessed or not assessed and single or multiple drafts. In addition, authors have used a number of different taxonomies, making comparison across different studies quite problematic. Inconsistent findings may also result from the rapid changes in exposure to technology both by the participants and by tools available to the researchers.

Writing process

Various models of the writing process have been put forward, which stem from the Hayes & Flower model (1980). This model viewed writing as a recursive and not a linear process (Weigle 2002:25). Van Waes & Schellens (2003:830) posit that it is implausible that one single writing process exists and suggest that cognitive processes are dependent on social and physical conditions and the writer's conception of the task. Their research showed that both individual characteristics and the mode of composition (physical environment) influenced writing processes. Burke & Cizek (2006:153) stress that writing is a thinking process that is guided by goals set by the writer; these goals are developed and modified throughout the writing process. Goals can be high-level, to do with meaning and direction or low-level, detail oriented such as spelling, capitalisation, punctuation and formatting. The authors hypothesise that a number of differences may be evident in compositions between administration modes due to high- and low-level goal setting and development. For example in PB composition, planning of text structure and order (setting high-level goals) is a critical step necessary before writing can commence due to the difficulty in making changes once the text has been generated. This initial goal making step is not necessary in CB composition due to the ease of revision and reordering.

Methodology

Research questions

The following research questions guided the study:

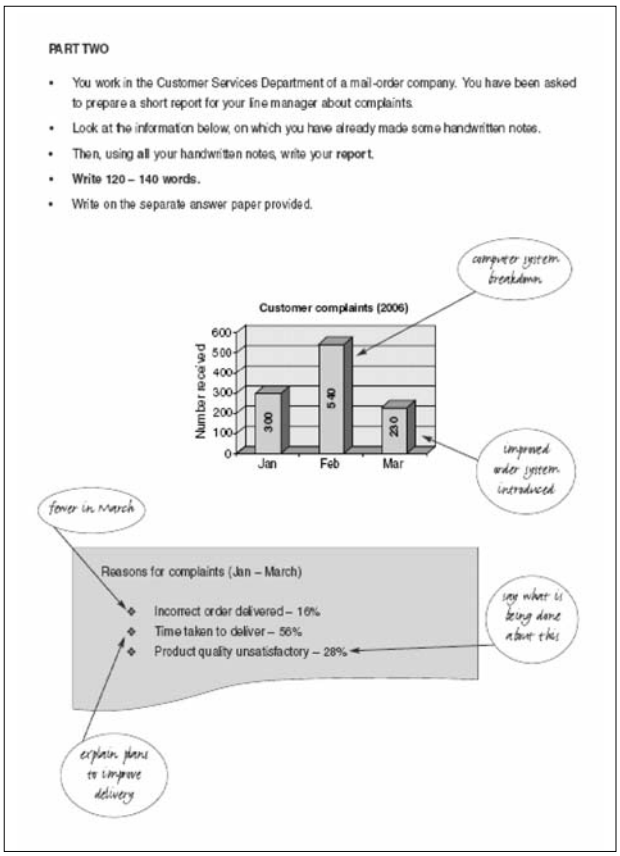
1) To what extent do candidates utilise the CB medium when writing a BEC Vantage Part 2 assessment task in terms of composition and revision strategies?
2) How do composition and revision strategies relate to writing score achieved?

Data collection and analysis

The focus of the study is Part 2 of the BEC Vantage Writing paper. BEC Vantage assesses English language ability used in a business context at Council of Europe 'Vantage' level (B2 on the Common European Framework of Reference – CEFR). It is available in a computer-based (CB) and paper-based (PB) format and consists of four papers: Reading, Writing, Listening and Speaking. In Part 2 of the Writing paper candidates are required to write either a piece of business correspondence, a report or a proposal. The composition is based on a rubric and input text(s) and should be between 120 and 140 words in length. The range of functions in the task may include explaining, apologising, reassuring, complaining, describing, summarising, recommending or persuading. The nature of the Part 2 task necessitates the need for an organised formal piece of text which should provide opportunity for candidates to demonstrate their composition and revision style. A sample Part 2 can be seen in Figure 1.

This study is an exploratory case study and as such is limited to 10 samples. Candidates from a single CB BEC Vantage session (December 2008) were sampled on the

Figure 1: Sample BEC Vantage Part 2 task



basis of exam grade (A–E) and, thus, their overall English language ability. A mix of passing (grades A, B and C) and failing (grades D and E) candidates was randomly selected. There were 10 candidates in total: three at grade C, two at grades A, B and D and one at grade E. All candidates answered the same questions. Eight of the sample came from the UK, and therefore represent a mix of L1s and two from a Mexican centre. Precise L1 data was not available for the sample. Five of the candidates sampled were male and five female. Ages ranged from 25–50, with a mean age of 31.

The data comprised of the response output obtained from the Connect test software which is referred to as a 'response history report'. The report contains a series of snapshots taken as the candidates compose their response. A snapshot is taken every 30–40 seconds and also each time a candidate moves between test parts. During the course of the Writing paper, test takers are able to switch back and forth between parts; this means that an individual could write their compositions across multiple and distinct blocks of time (henceforth referred to as sub-parts). Thus from the response history report we are able to establish in which order candidates have completed the two writing parts, whether they have switched back and forth between parts and how long they have spent on each part. Within the response history output, times are only recorded in whole minutes so any times given in this paper are approximate.

The data were manually classified into four categories: unchanged, new text, low-level revisions and high-level revisions.

Unchanged was used if the composition did not change from the preceding snapshot.

New was used if text was added at the end of the composition. The end was stipulated as this is something that could easily be done in both PB and CB compositions.

Insertions of new text mid-composition could only easily be achieved in CB mode and were counted as a revision. Candidates sometimes make changes to the last few words of the text as they write. It was decided not to class these as revisions, as the time interval of snapshots would not enable accurate coding.

All other changes made to the text were classed as revisions. Revisions were grouped into low- and high-level changes following Li (2006). Low-level revisions are at or below the word level and high-level revisions are at or above the phrase level. When composing on paper it is more difficult for candidates to make high-level changes within existing text so it was considered important to assess the extent and frequency of different kinds of revision. Low- and high-level were split into sub-categories (see below) so as to ascertain more detail about the types of revision made.

Low-level revisions:

- insert word
- delete word
- correct a spelling
- punctuation (insert/delete/change)
- word edit (e.g. played to play)
- word change/lexical substitution (e.g. aim to purpose).

High-level revisions:

- insert phrase
- insert sentence
- delete phrase
- delete sentence
- re-write phrase
- re-write sentence
- cut and paste at or above the phrase level.

Using the coding output from the Part 2 task, the author mapped activity (unchanged, new text, low-level revisions and high-level revisions) across the time duration of each composition. In addition, frequencies of each activity were counted so that comparisons could be made about prevalence of each activity and the relation between activity and test score.

Results and discussion

This section investigates the stage at which revision and composition activities occur within the writing process and the duration of the activities; it then shows how these relate to score. The section concludes with a focus on the frequency and nature of revisions.

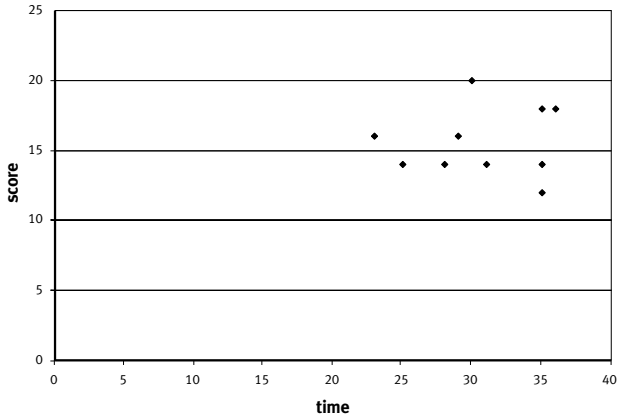
Firstly, the amount of time spent on the task was analysed. This was done to establish whether there was any time effect present that could impact on any conclusions concerning activity and score.

Table 1: Part 2 score obtained by each candidate

Candidate	A	B	C	D	E	F	G	H	I	J
Score	14	18	14	14	16	18	20	12	16	14

The findings showed that time spent on Part 2 varied from 23 to 36 minutes, the average time being 31 minutes (SD =4.5). Part 2 scores ranged from 12 to 20 (out of 20), with a mean of 15.6 (see Table 1). When candidates' Part 2 times were compared to their Part 2 Writing score, no pattern emerged (see Figure 2): this indicates that spending more time on Part 2 does not necessarily result in a higher score.

Figure 2: Comparison of response times and scores for Part 2



Revision and composition activities undertaken with reference to time and stage

Figure 3 shows the stage at which activities occur throughout the composition and their duration. Sub-parts, distinct composition blocks as a result of switching between parts, are indicated by a bold horizontal line. In sections where both low- and high-level revisions occur, the high-level ones are recorded as these were the dominant activity.

If we first look at the number of sub-parts by candidate, we can see that the majority of candidates switched back and forth between test parts in the course of the test. There were in fact more switches than indicated on Figure 3, but as their duration was less than one minute they were too short to be recorded.

The freedom to switch between parts reflects the format of the PB test. The fact that seven out of the 10 of the CB candidates used the switching facility provides support for the inclusion of this feature in the CB format. That the two modes share this feature adds strength to the comparability of their test scores. In addition, writers are able to switch from task to task in the target language use (TLU) domain, so it is important that the test reflects this for the reasons of context validity.

As far as overall composition strategy is concerned, nine of the 10 candidates essentially composed from beginning to end, although they did return to earlier portions of the text to make revisions. One candidate (F) wrote key words and a skeleton structure and then used this to build up the composition:

'Firstly, Summary the aim of this report is to assess the possibility of attending the European Trade Fair in London. Findings Overall, Firstly, In terms of, Furthermore, Conclusion, I therefore suggest to attend the European Trade Fair in May'.

This skeleton structure was the sole incidence of onscreen planning behaviour, although it was not of the content-based kind. Candidates B and G did have sub-parts with no recorded writing activity which could have been used for mental planning. The Cambridge ESOL Teacher Support website clearly states in the advice for teachers section for BEC Vantage to 'train students to plan before they write', so this lack of planning evidence could be disquieting and could also support Li (2006) who noted that writers engaged in less planning when composing with a word processor. However, it could be argued that the BEC task itself provides a lot of structure, which is why formal planning was not deemed necessary. Moreover, it should be noted that although candidates are not provided with paper and pencil, they do have access to a piece of paper with their candidate token on, so it is possible that this was used for planning activities.

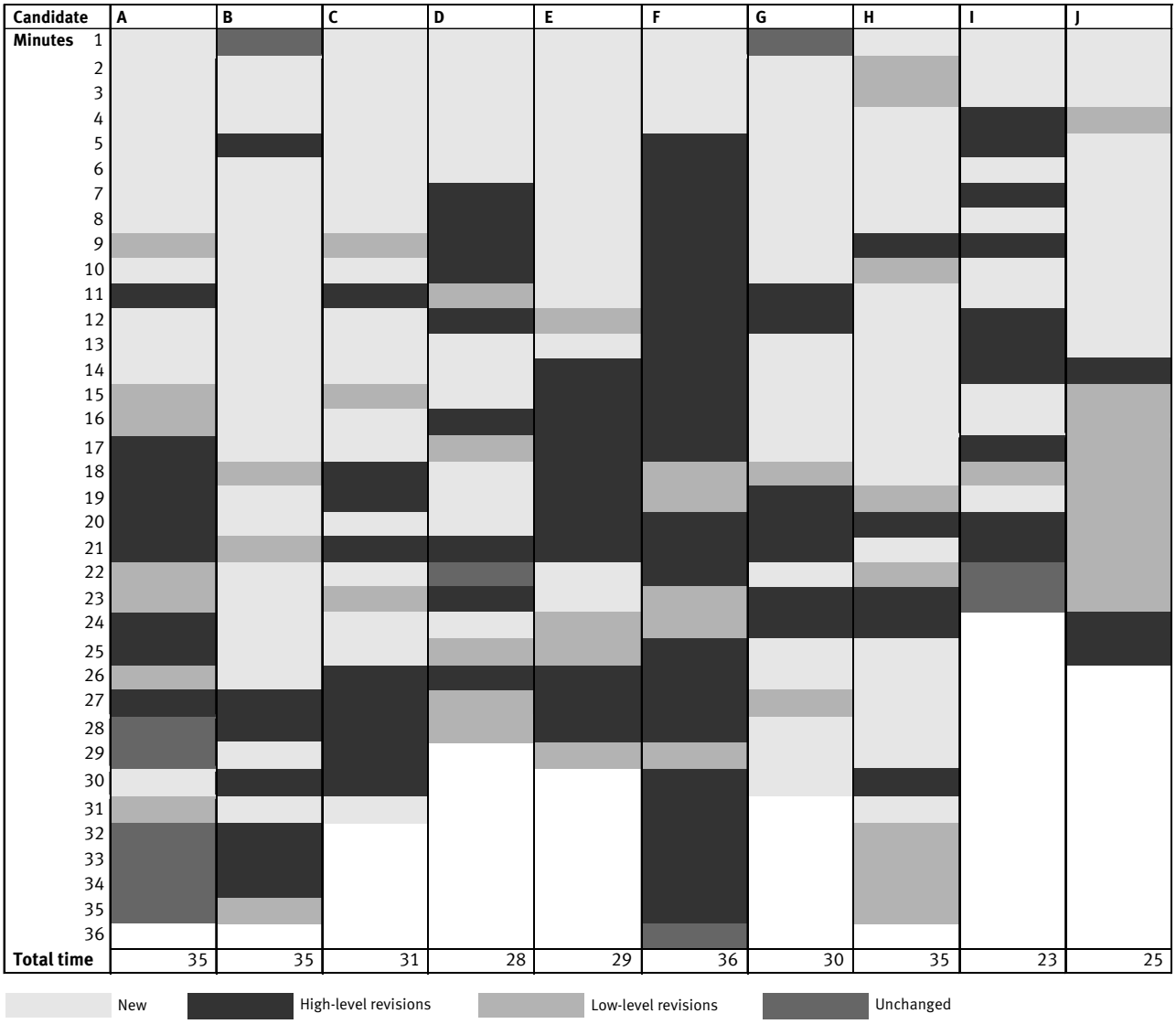
All the candidates in this study did engage extensively in modifying the last few words of text as they typed new text, which may suggest that composition was 'type then monitor' rather than 'plan then type'. This fits with the hypothesis of Burke & Cizek (2006), discussed above, who assert that goal setting is not necessary in CB composition due to the ease of revision and reordering. Lee (2002:152) also notes that planning and text production on the computer appear to be more interwoven than they are on paper.

It is perhaps surprising that only one candidate used a skeleton structure and infill approach. It shows that, in general, the candidates did not fully utilise the medium in terms of the flexibility of composition approach that word processing allows. Instead the candidates composed in a similar way to how one would on paper (from beginning to end). This could be the influence of it being a timed assessment: candidates may go into 'exam mode' concentrating on getting their compositions typed.

In terms of production and revision activities during the composition, with the exception of candidate F described above, four candidates wrote their text and then revised it (A, B, E & J), another four candidates wrote text and then engaged in a mixture of revision and additional new text (C, D, H & I) and one candidate wrote new text, revised it and then wrote more new text (G) (see Figure 3). Composition strategies appear to depend very much on the individual and their interaction with the administration mode, with some candidates having distinct periods of composition and revision and some switching repeatedly back and forth. When related to score, none of these strategies produced distinctly higher scores and it should be remembered that L2 proficiency also contributes to marks. Revision activities will be further related to score in the next section.

As can be seen from Figure 3, the balance of time spent composing new text and being engaged in revision varied between candidates. With the exception of candidate F who used the infill approach, over 40% of the time was devoted to the production of new text. The highest scoring candidates spent at least 70% of the time in text

Figure 3: Stage and time engaged in composition and revision activities



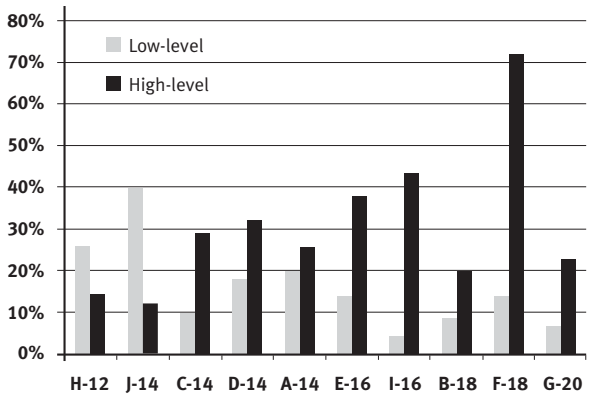
production. It is interesting to note that both these candidates (B and G) wrote nothing in the first sub-part, perhaps indicating that some mental planning was taking place. However, two of the lower scoring candidates, H and C, also spent a large proportion of their time (at least 60%) on new text production. This would appear to indicate that a higher proportion of time spent engaging in new text production does not necessarily result in a higher score.

There were few time periods where no change occurred to the composition. This shows that virtually all the time was spent actively engaging in either composing or revising text. The fact that this composition was timed would probably account for this.

Revision and composition activities undertaken with reference to time and score

All candidates engaged in text revision. When proportion of time spent on high- and low-level revisions is examined with respect to score (see Figure 4), it can be seen that, in general, candidates who attained lower scores spent more time making low-level revisions and less time making high-level ones. The opposite applies to higher scorers. In one sense, this supports the literature on experienced/inexperienced writers revising differently (Faigley & Witte 1981, Somers 1980). However, it must be remembered that in this context scores will also reflect L2 proficiency in addition to writing expertise.

Figure 4: Proportion of time spent on different revision activities (candidates in score order)



Candidate I, who spent the greatest proportion of time making high-level revisions (besides candidate F), spent a lot of time reworking their composition, inserting, deleting and re-writing sections. However, their score was mid-range indicating that it is quality and not amount of revision that is important. The candidate who achieved full marks spent the least amount of (actual) time revising (candidate G). Plakans (2008) stresses the importance of the contribution of L1 writing skill to L2 writing. Based on that, it could be hypothesised that candidate G is an experienced/expert L1 writer, and that these skills have transferred to L2 writing in this case. Possible evidence of this is that this candidate's overall exam grade was C and, yet, the candidate scored full marks in this task.

Frequency and nature of revisions

Figures 5 and 6 show a breakdown of revision type summarised for the study participants as a whole. It can be seen from Figure 5 that low-level changes concentrated on inserting/deleting/replacing a word rather than on 'proofing' changes such as spelling, capitalisation and punctuation. This could reflect an over-reliance on the auto-correct features and spellcheckers that accompany word processing packages; users may pay less attention to proofing activities as the software does it for them. Crawford et al (2008) also found a similar pattern although they found a relatively higher incidence of spelling changes. The author felt that all compositions could have benefited from proofing refinements and it is recommended that candidates/teachers bear this is mind.

When looking at high-level changes (see Figure 6), it is interesting to note that cutting and pasting did not appear to be used often, though the author acknowledges that due to the time intervals at which snapshots were taken it was often difficult to code accurately. The focus seemed to be more at phrase level changes than at full sentence level, a finding similar to Crawford et al (2008) who found that the most frequent activity was inserting a phrase.

We can look further at the revision activities investigated and explore whether these are a facet of the CB mode. It can be safely assumed that all low-level changes could be made in PB mode in addition to CB mode. If pencil is used, these can be made relatively easily; a pen would have a more untidy result, but low-level changes would still be feasible. With the high-level changes, the author believes it is also possible to split these into activities that could and could not be achieved in PB mode. It is hypothesised that inserting/deleting a phrase/sentence is possible and that re-writing a phrase/sentence and cutting and pasting are not possible in a paper-based mode. If this is taken to be the case then Table 2 shows the proportion of activities available in PB mode. It can be seen that between 73% and 100% of revision activities would be possible. However, in PB mode, if a candidate were to make the number of revisions made in these examples, the script would be very messy. It appears to be the opportunity to make frequent revisions rather than the nature of the revisions *per se*, that these candidates have utilised in the CB mode.

Figure 5: Comparison of the use of low-level revision types

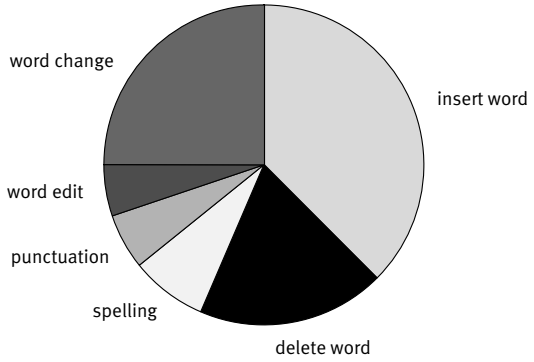
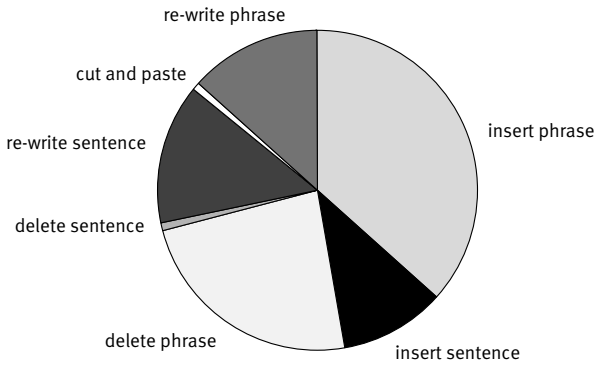


Figure 6: Comparison of the use of high-level revision types



Interestingly, the candidate with the lowest score (H) had the highest proportion of activities not possible in the PB mode. This candidate seemed to be spending time reworking parts of the text without improving the content of the composition; this again highlights the importance of quality of revision. Pre-planning might have mitigated the need for this.

Conclusion and recommendations

Candidates used a variety of composition/revision strategies, which appear to depend very much on the individual and their interaction with the administration mode, with some candidates having distinct periods of

Table 2: Proportion of revisions which are possible and impossible in PB mode

Candidate	A	B	C	D	E	F	G	H	I	J
Revision activities										
PB possible changes	85%	73%	85%	85%	94%	79%	86%	72%	85%	100%
CB only changes	15%	27%	15%	15%	6%	21%	14%	28%	15%	0%

composition and revision and some switching repeatedly back and forth. The research questions and their answers are summarised below.

To what extent do candidates utilise the CB medium when writing a BEC Vantage Part 2 assessment task in terms of composition and revision strategies? It is evident that candidates did not fully utilise the CB medium in terms of the flexibility of composition approach that word processing allows and composed in a similar way to one would in a PB test (i.e. from beginning to end). This could well be an effect of it being a timed assessment, in that candidates want to get their ideas/text written down quickly. Also, it could be a result of the scaffolding structure and the level of cognitive challenge required by the task. There was little evidence of onscreen planning, but it is possible that some planning may have been done on a piece of paper. It is recommended that teachers/candidates heed the advice for planning their composition in either PB or CB mode and to teach CB candidates how the mode could be used to aid planning.

Candidates engaged in both low- and high-level revision activities, with the proportion varying between candidates. Low-level changes mostly involved inserting/deleting/replacing a word and the high-level changes were mostly at the phrase rather than sentence level and usually involved inserting a phrase. This supports the findings from the previous literature which suggests that CB revisions tend to focus on lower-level linguistic features (Crawford et al 2008:112, Van Waes & Schellens 2003:348). There were limited proofing changes, such as spelling and punctuation, although all compositions could have benefited from them; it is recommended that the importance of proofing is emphasised in writing classrooms. The majority of the revisions made could have been made in PB mode; however, it could be argued that the frequency at which they occur would be unlikely in a PB script due to the resulting untidy appearance of insertions, crossings out, etc.

The study shows that candidates do utilise the CB administration mode when revising, but the frequency of the CB-only activities is not high. It appears that what the candidates utilised in the CB mode was the opportunity to make frequent revisions rather than the nature of the revisions *per se*. This could be a facet of the task itself in terms of scaffolding, time and length. These limited findings add strength to the comparability of PB and CB timed written assessments as CB candidates do not appear to exhibit different behaviour to PB candidates.

How do composition and revision strategies relate to writing score achieved? No relationship was found between composition strategy and score in this limited sample. Time spent on revision activities themselves changed with candidate. In general, candidates who attained lower scores spent more time on low-level revisions and less on high-level ones, while the opposite is true for high scorers. The revision activities carried out do not necessarily lead to better scores. This highlights the importance of quality of revision. It is recommended that teachers should advise students about using revision appropriately, i.e. tailoring the revision strategy to the type of composition and its constraints and not focusing too much on revising because one feels one should or because the mode allows it. One of the limitations of this kind of analysis is the fact that composition/revision is examined only from the perspective of the compositions. An improvement would be to couple this analysis with candidate interviews, using the response history report as a basis for collecting candidate perceptions on how they were composing and more importantly why they were composing in this manner.

The task used in the study was not overly challenging and scaffolding was provided, which further decreased the challenge. It would be interesting to conduct a similar study on higher proficiency levels to investigate whether a more cognitively and linguistically challenging task would produce similar findings. In addition, investigating which specific revision behaviours result in improvements to text quality would aid the understanding of which revisions to attend to in timed writing.

It could be argued that the timed, assessment focus of this study is too narrow and that it does not fully capture the writing process. However, time-constrained writing is a facet of real life, especially in the business world. It is important that students learn how to compose and revise effectively in such conditions, if they are to use L2 successfully at their work or place of study. This study suggests some aspects for teachers and candidates to focus on. Ultimately, it is important that students learn how to revise in a way suitable for a given context. Worden (2009:176) is undoubtedly right in arguing that 'those students that have at their disposal a variety of composing strategies and are able to critically reflect on the requirements of the specific writing context to choose among them will be the best prepared'.

References

Burke, J and Cizek, G (2006) Effects of composition mode and self-perceived computer skills on essay scores of sixth graders, *Assessing Writing* 11 (3), 148–166.

Crawford L, Lloyd, S and Knoth, K (2008) Analysis of student revisions on a state writing test, *Assessment for Effective Intervention* 33 (2), 108–119. Retrieved on 12/07/10 from http://aei.sagepub.com/content/33/2/108.short

Faigley and Witte, S (1981) Analyzing revision, *College Composition and Communication* 32 (4), 400–414. Retrieved on 29/06/10 from http://www.jstor.org/stable/356602?origin=crossref

Hayes, J R and Flower, L (1980) Identifying organization of writing process, in Gregg, L and Steinberg, E (Eds) *Cognitive processes in writing*, New Jersey: Lawrence Erlbaum Associates, 3–30.

Lee, Y (2002) A comparison of composing processes and written products in timed-essay test across paper-and-pencil and computer modes, *Assessing Writing* 8 (2), 135–157.

Li, J (2006) The mediation of technology in ESL writing and its implications for writing assessment, *Assessing Writing* 11 (1), 5–21.

Plakans, L (2008) Comparing composing processes in writing-only and reading-to-write test tasks, *Assessing Writing* 13 (2), 111–129.

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

Somers, N (1980) Revision strategies of student writers and experienced adult writers, *College composition and communication* 31 (4), 378–388.

Stevenson, M, Schoonen, R and de Glopper, K (2006) Revising in two languages: A multi-dimensional comparison of online writing revisions in L1 and FL, *Journal of Second Language Writing* 15, 210–233.

Van Waes, L and Schellens, P J (2003) Writing profiles: The effect of the writing mode on pausing and revision patterns of experienced writers, *Journal of pragmatics* 35 (6), 829–853.

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

Worden, D (2009) Finding process in product: Prewriting and revision in timed essay responses, *Assessing Writing* 14 (3), 157–177.

Effective pretesting: An online solution

LAURA COPE RESEARCH AND VALIDATION GROUP, CAMBRIDGE ESOL
ANDREW SOMERS RESEARCH AND VALIDATION GROUP, CAMBRIDGE ESOL

Introduction

As part of Cambridge ESOL's routine test production cycle, all test material passes through the pretesting process. Pretesting is a key stage in the overall quality control processes that ensure we can deliver valid and reliable examinations. In essence, the pretesting process is a trial of future test material on a sample of candidates. Subsequent qualitative and quantitative analysis of the data collected from the pretests is then used to select the most suitable material for use in live examinations, or to refine items which did not meet Cambridge ESOL's test construction standards. Moreover, the analysis carried out during the pretesting process enables us to ensure all versions of our examinations are constructed to consistent levels of difficulty, do not unfairly differentiate between certain groups of candidates, and are effective and reliable instruments for measuring candidates' language proficiency.

This article begins with an overview of Cambridge ESOL's traditional paper-based (PB) pretesting process and some of the challenges it must overcome to ensure its effectiveness. We illustrate how this approach is designed to meet the theoretical requirements to provide a reliable way of equating tasks within our item banking system, which in turn enables us to consistently deliver exams at the appropriate levels. (For more information on our item banking system see Beeston 2000.) Balanced against these are practical limitations on the volume of pretesting that can be carried out. We then move on to discuss how pretesting can be delivered through the use of new technology with an Online Pretesting System. We explain how this system operates in the context of adaptive testing, and how this leads to a more efficient and flexible process, when compared to the more traditional PB approach, while also maintaining the accuracy and security of the methods used for analysing the data and calibrating material.

Challenges in paper-based pretesting

The current paper-based pretesting approach is a considerable undertaking spread over several months involving many resources: in the last year 1,200 pretests were administered to 120,000 candidates. Before analysis and review can even take place, the administrative process at Cambridge ESOL involves initial invitations to centres to recruit candidates, allocating a range of pretests to the appropriate candidates, and the subsequent marking and capturing of candidates' responses and marks. Once collated, the data is then subject to Classical and Item Response Theory analyses to provide statistical information on the performance of items. In parallel, the individual responses to items are also collated to enable a qualitative view of how candidates perform on certain tasks. Both views are then considered when reviewing each item/task and assessing their suitability for use in the final examination.

The key to the whole pretesting process is to trial the test material on an appropriate population to ensure we get a clear picture of how our tasks will eventually perform in the live examination. Cambridge ESOL's current paper-based pretesting model targets candidates who are preparing to take a Cambridge ESOL examination in the near future – usually in the final weeks before they take the exam. We also provide pretests which are as close as possible in format to the live exam, and require the conditions in which they are taken to be equivalent to those on the examination day itself. By targeting a sample of our actual live candidature, working under the same conditions as a live examination and with similar preparation and familiarity with the examination content and structure, we are able to ensure that the pretest candidates are as close as possible to the actual live population. This approach also provides a valuable opportunity for the candidates to practise, and receive feedback on, examples of the real examination, enabling us to find willing volunteers with an incentive to partake in the pretesting exercise.

A variety of personal characteristics, henceforth referred to as Candidates' Information Sheet (CIS), are monitored in the pretesting population, to ensure that we do indeed pretest material on a representative sample of candidates. The CIS includes features such as age, gender and first language. Considering these characteristics helps to ensure that our test material does not exhibit any bias towards any particular subgroup of candidates, and remains a test of language proficiency in the relevant context. Furthermore, to achieve the maximum statistical information about the performance of items, the pretest candidature needs to be of a similar level of proficiency to that which the tasks demand. Clearly these features of the candidates are unknown during the time of recruiting the candidates. We thus target a range of Cambridge ESOL centres around the

world, with the expectation that through their knowledge of the tests, we will secure a sufficient number of candidates exhibiting the desired features. Nonetheless, one cannot always get a perfect group of candidates, which is why some further tailoring of the sample is required to achieve the optimal results. On occasion, it has been necessary to require top-ups of pretests, where the original sample was insufficient to meet the appropriate CIS criteria, such as a spread of different first languages. Topping-up a pretest with further data requires an additional cycle of recruiting candidates and administering the tests, which requires significant further resources.

Once an appropriate sample has been obtained, the analysis of item performance, and in particular the determination of task difficulty, presents further challenges to successful pretesting. Cambridge ESOL has created a Common Scale within its item banking system which covers the full range of Common European Framework of Reference (CEFR) levels and all Cambridge ESOL exams (see Cambridge ESOL website). All pretested tasks are anchored to the Cambridge ESOL Common Scale through anchor tasks, which is of paramount importance in enabling us to produce examinations at an appropriate and consistent level of difficulty.

To ensure a suitable degree of accuracy in the calibration of items – and to locate them on the Common Scale – we require the anchor tasks to meet certain criteria. The absolute number of anchor items and their proportion relative to the pretest items which are to be calibrated is the first concern. Too few anchor items and there is insufficient data on which to undertake an accurate analysis. Too many anchor items can render the pretests too long and introduce undesirable side-effects such as tiredness. Alternatively, if the overall test length is preserved, using too many anchor items will limit the number of new items that can be evaluated.

Recommendations for a suitable number and proportion of anchor items typically suggest around 20–30% of the total number of items in the test and a minimum of 20 items are sufficient for most purposes (Angoff 1971, Kolen & Brennan 2004). However, in a 25 item Listening test for example, there is a challenge to provide sufficient anchor material, pretest enough new material and maintain a reasonable test length that is representative of the intended examination. Anchor items also need to be representative of the tasks being calibrated in terms of content, such that performance on one task is a sufficiently valid and reliable indicator of performance on other tasks. The material also needs to be suitably targeted at the appropriate level of difficulty to ensure that measurement errors are kept to a minimum acceptable level.

Thus, in order to carry out comprehensive pretesting of material to meet the requirements and deliver numerous test versions each year, our model has evolved to balance the competing requirements for our needs. However, there are potentially other ways to address these problems, which in turn create different challenges. For example, the difficulties in achieving a suitable sample of candidates could be overcome through the use of pretesting material in live tests. This would enable access to a much larger and more diverse pool of candidates, and allow pretesting to take place, whilst exactly replicating live test conditions. However, doing so and generating the required volume of material would involve the production of many more differing test versions than we currently do. Moreover, in exams which have a clearly defined fixed format, pretesting material could not necessarily be additional to the test, but would need to replace part of the live exam. In such cases, there is an issue surrounding the grading of candidates, equating the multiple versions and whether it is fair to administer untrialled material in this way on an unsuspecting candidature. Some of these possibilities are explored and addressed through the online pretesting model which we present here.

Online pretesting

Advancements in technology are enabling our current pretesting methodologies to be reconsidered and improved. For example, pretesting items within an online environment has the potential to solve many of the challenges associated with traditional paper-based pretesting discussed previously. There are various approaches to carrying out online pretesting, but this paper focuses on the benefits of online pretesting within the context of computer-adaptive tests (CAT).

Each candidate who takes a CAT test receives a unique set of items. The test adapts the difficulty level of the items that it administers to each specific candidate, based upon the candidate's estimated ability throughout the test, resulting in individual tailor-made tests. Compared to traditional paper-based linear tests, CAT tests require fewer items to obtain scores of equal accuracy (Weiss & Kingsbury 1984). This set-up provides excellent opportunities for carrying out the pretesting of new material through the automated test-assembly process and the flexibility that the assessment design allows.

Cambridge ESOL has recently developed the functionality to carry out online pretesting within CAT tests. The new items which have not yet been pretested and hence equated within our item banking system are known as uncalibrated items; calibrated items have associated difficulty values obtained via the pretesting process. Uncalibrated items, which undergo a series of editing and quality assurance cycles before selection for pretesting, are assigned a provisional difficulty estimate based upon expert examiner judgement (examiners classify items as low, medium and high difficulty) and sit within the same item bank as live calibrated items. During item selection within a candidate's test, the pretest items and live calibrated items are treated equally until the limits, which are imposed by the CAT algorithm upon the maximum numbers of uncalibrated items that the tests may administer, are exceeded. These limits ensure that the proportion of uncalibrated material within tests is small. The uncalibrated items are embedded within the live candidate tests and are selected in a randomised nature according to the specified CAT algorithm. At the end of the test, the candidate's responses to any uncalibrated items that they received are excluded from the calculation of their estimated ability and result, thus eliminating any unfairness

in test results due to exposure of untrialled material. By incorporating uncalibrated items within the live CAT tests, the need to produce separate standalone tests composed of uncalibrated items (pretests) is eliminated. This removes many unnecessary procedures and paperwork surrounding traditional paper-based pretesting, such as composing these tests and organising centres to administer them, thus improving the efficiency of the administration. The flexible nature of the CAT test allows the length and composition of the test to vary within certain parameters enabling us to add in additional items for pretesting, without impacting on the validity of the candidate's result.

There is also significantly more flexibility in when and how items are pretested in CAT tests. Potentially any number of items can be pretested at any one time, from a single item through to many hundreds of items. Items can begin being pretested whilst other items are still in the middle of the pretest process. This is all due to the flexibility of the system to allow items to be added and removed from the online item bank whenever we wish. Additionally, if an item is amended following pretesting, it can easily be re-pretested without needing to be incorporated into a new linear pretest. The pretesting process is no longer a series of discrete administrations and analyses, but an ongoing continuous process allowing items to be pretested as much and as often as necessary to achieve sufficient information.

CAT tests are suitable for candidates with a wide range of abilities since they adapt to the ability level of each individual candidate. Consequently, candidates from across the full spectrum of abilities take the same CAT test. To cater for these needs, the item bank must contain items from across all difficulty levels. This in turn allows the pretesting of material from the full range of difficulties within a single CAT test. Furthermore, the pretesting could be targeted at only a very small subset of difficulties (for instance very high difficulty items) where new material is most desired. Achieving this with traditional paper-based pretesting would be very difficult.

When pretesting an item, maximum information about the item's difficulty is gained when the proportion of correct candidate responses (facility) is 50%. In a paper-based linear pretest, this is not practical to achieve since each item will be administered to candidates with a range of abilities. However in a CAT test, the item's provisional difficulty estimate forces the item to be targeted at candidates with ability close to this figure. The item's facility during the pretesting phase can be monitored and the item's assigned difficulty estimate can be adjusted, if necessary, to bring the facility closer to 50%, resulting in more information about the real item difficulty.

The fact that the pretest items are targeted at a much narrower and appropriate range of candidate abilities means that information about the item difficulty is gathered more efficiently than in paper-based pretesting. Therefore, fewer candidates per item are required for calibration purposes. This speeds up the pretesting cycle and is, therefore, an important benefit of online pretesting.

When pretesting an item, it is important that the item is administered to a sample of candidates with a range of first language (L1) backgrounds, to prevent the item from being calibrated with a bias towards a particular L1 group. For example, if a new item is administered to a group of candidates, 90% of whom are native German speakers, the item's difficulty estimate may be biased if a particular aspect of the German language enables these candidates to respond to the item more successfully than candidates from other backgrounds. Online pretesting through the CAT test has the advantage of using the live examination population for pretesting material – thus the entire population is available from which a sample of candidates is presented with pretest material. This, coupled with the randomised nature of item administration, ensures that each item is seen by a healthy mix of candidates and corresponding first languages. Online pretesting within CAT tests provides an ideal solution to this requirement, which is often problematic in paper-based pretesting where the pretests are taken by only a few centres.

The whole process of pretesting online is made very efficient by having all data in a computer-based format. There is no need to spend time transferring candidates' responses into a computer and all data is easily available and ready for analysis.

The calibration of items that have been pretested online within CAT tests is very accurate. The live calibrated items and pretest items that are randomly administered within candidate tests are not distinguishable by candidates. This creates an ideal situation for the pretest items to be administered to candidates who will respond to them with the same level of motivation as they would to calibrated items. Hence the items' calibration results are a direct result of candidates' abilities and do not incorporate candidate motivation levels.

There is no such thing in a CAT test as not getting to the end of the test and leaving the last few questions unanswered. Every item that a candidate receives in a CAT test will not be removed from the screen until the candidate responds. Therefore any pretest items that are included in the test will be seen by the candidate and will be responded to, eliminating from the pretesting calibration any incorrect candidate responses due to candidates progressing too slowly through the test and not reaching the final questions.

The pretesting analysis for item calibration works fundamentally in the same way as in paper-based pretesting, where a set of 'anchor' items is used, with known, trusted difficulty values, from which the pretest items are calibrated. Each previously calibrated and each new pretest item in the online item bank is administered to a unique set of candidates, which results in a comprehensive linking network between candidates and items. A link between two items can occur if both items are taken by a common candidate. This vast linkage within the item bank means that the whole set of calibrated items can potentially be used as anchor items, leading to a high level of accuracy when determining the difficulty estimates of the new pretest items.

CAT tests are inherently very secure due to the large size of the underlying item banks and the random distribution of items across test centres and countries. This in turn leads to a good level of security for pretesting within CAT tests. A candidate who memorises a few isolated items from their CAT test does not pose significant risk to the security of the

material or the bank. Additionally, online pretesting removes the need for paper copies of materials being shipped around the world.

Ever since online pretesting has been implemented within CAT tests at Cambridge ESOL, a few hundred items have been pretested within a short and very limited time period. Using paper-based pretesting, this same feat would have taken months to achieve with many more candidates. Online pretesting has the potential to automatically access a large number of candidates as opposed to manually inviting and recruiting candidates centre by centre. Thus it has the potential to dramatically improve the efficiency of the pretesting process. Current paper-based pretests are administered to far fewer candidates who take our live exams whilst online pretesting has the opportunity to reach every candidate if necessary, and present different items to each one of them. There is, therefore, the potential for online pretesting to significantly increase the volume of material which we could pretest, and do so in a more efficient and accurate manner.

References

Angoff, W H (1971) Scales, norms and equivalent scores, in Thorndike R L (Ed.), *Educational Measurement* (2nd Ed.) Washington DC: American Council on Education, 508–600.

Beeston, S (2000) The EFL Local Item Banking System, *Research Notes* 1, 5–6.

Cambridge ESOL, retrieved from http://www.CambridgeESOL.org/what-we-do/research/cefr/item-banking.html

Kolen, M J and Brennan, R L (2004) *Test Equating, Scaling, and Linking Methods and Practices*, New York: Springer.

Weiss, D J and Kingsbury, G G (1984) Application of computerized adaptive testing to educational problems, *Journal of Educational Measurement* 21, 361–375.

Conferences and publications

HR Magazine conference

HR Magazine organised a one-day conference for HR managers in Hong Kong on 22 July 2010 where over 250 HR managers from various Hong Kong based companies attended. Cambridge ESOL China Office sponsored the afternoon session: 'Innovation in staff benchmarking – measuring training effectiveness'.

Dr Ardeshir Geranpayeh delivered the plenary lecture in the afternoon on 'Benchmarking language proficiency for the workplace'.

In his presentation, Ardeshir discussed the need for improving language proficiency in the workplace. He started by linking the global success of international companies to effective knowledge transfer which has roots in increasing remote management, which in turn requires companies to work from multiple locations. He went on to say that in such a community individuals bring their own cultural and linguistic background into the workplace, which makes virtual communication both internally and externally a challenging task. This is why improving language proficiency in the workplace becomes an important policy for international companies. To achieve that, one needs to benchmark jobs to language proficiency requirements. He discussed examples of successful benchmarking that Cambridge ESOL has undertaken so far and outlined the features of a benchmarking tool Cambridge ESOL has developed. He reported on successful use of such a toolkit in airline, recruitment, banking and financial organisations.

BAAL Testing, Evaluation and Assessment SIG conference

The BAAL TEA SIG held a half-day conference at the Centre for English Language Education (CELE) at the University of Nottingham in November 2010, with the theme of 'Language tests for immigration: Conflicting Ideologies and Challenges'. The current and topical theme of the conference attracted delegates from a range of UK institutions. The programme included an Introduction by Vivien Berry, the SIG Convenor, followed by presentations on language assessment policies in the UK and Europe. In the first presentation, Diane Schmitt (the Testing Officer for BALEAP) provided an overview of the current guidance from the UK Border Agency for Tiers 1, 2 and 4, giving special attention to the sections which describe English language proficiency requirements and the type of evidence that will be accepted as proof of English language proficiency. Marli Tijssen, from the Centre for Innovation of Education and Training (CINOP) in the Netherlands, overviewed the immigration and language policies in the Netherlands, including a 2006 law on the integration of immigrants, which stipulates that spoken language and culture exams in the immigrant's home country are compulsory before arrival in the Netherlands.

The conference also included the test providers' perspectives, with presentations from Lee Knapp (Cambridge ESOL), Alan Baldock (Password Partnerships), and David Booth (Pearson Language Tests). Lee Knapp argued for the importance of a system which is appropriate, fit for purpose, reliable and secure, and emphasised Cambridge ESOL's commitment to supporting the development of a language and migration policy and system that works for all. Alan Baldock focused on the assessment of incoming international students by universities and the role of the Password test as an assessment of students' English language level. David Booth discussed the need for secure test delivery and demonstrated how the Pearson Test of English (PTE) Academic responds to such challenges. The conference ended with a summary of the issues by Professor Barry O'Sullivan (Roehampton University), and a roundtable discussion.

English Profile events

Autumn 2010 saw a series of events and talks on the English Profile (EP) around Europe. Starting in Madrid, Spain, the EP team, including Cambridge ESOL's Prof. Roger Hawkey and Dr Angeliki Salamoura, held two English Profile seminars at the Colegio Oficial de Doctores y Licenciados en Filosofia Letras y en Ciencias on 22 and 23 October. On the first day, the talks concentrated on in-depth research findings of the English Profile Programme, whereas the talks on the second day were shorter and aimed at teachers.

On 19 November, Dr Angeliki Salamoura and Dr Nick Saville gave a presentation entitled 'Exemplifying the CEFR: Findings from the EP Programme' at TESOL-Italy's XXXV National Convention in Rome.

Finally, Dr Angeliki Salamoura, Dr Julia Harrison of Cambridge University Press and Milan Milanović of the University of Kragujevac, Serbia, led an EU-funded English Profile workshop on 'The CEFR and English Profile in the classroom' across three locations in Serbia in November: Megatrend University, Belgrade; the University of Kragujevac, Kragujevac, and the International University of Novi Pazar, Novi Pazar. Angeliki led practical sessions aimed at familiarising teachers with the CEFR and the EP aims as well as the latest findings. Julia talked about the Cambridge EP Corpus as a classroom resource and Milan provided a practical demonstration of the data entry process on the English Profile Data Collection Portal. For more information on the English Profile Project, please visit www.englishprofile.org

ALTE events

ALTE has recently completed a successful week of activities in Prague centring on its biannual meeting and conference. The week began with a 2-day Extended Learning Course on 'The Application of Structural Equation Modelling (SEM) in Language Testing Research' which was run by Dr Ardeshir Geranpayeh, Assistant Director, Research and Validation, Cambridge ESOL. Participants came from several countries including the Czech Republic, Poland, Norway, France and Portugal.

The ALTE meeting and conference took place from 10–12 November. The conference theme – 'Fairness and Quality Management in Language Testing' – reflects one of ALTE's key objectives, namely to maintain common standards throughout all stages of the language-testing process. During the course of the workshops and presentations the presenters looked at ways of ensuring the fairness of language tests and testing practices, and of ensuring the quality demands of examinations in relation to their functions and purposes.

In September ALTE ran two very successful summer testing courses in Bilbao. These courses were hosted by ALTE's Basque member, the Basque Government, and attracted participants from all over the world. Lynda Taylor and Cyril Weir ran the first week's course – 'The ALTE Introductory Course in Language Testing' – and Ivana Vidaković and Angeliki Salamoura ran the second week's course – 'The ALTE Introductory Course in Testing Reading'.

Looking ahead to 2011, the ALTE 4th International Conference will take place from 7–9 July at the Jagiellonian University in Kraków, Poland, and online registration is now open – http://www.alte.org/2011/registration.htm

The theme of the conference is 'The Impact of Language Frameworks on Assessment, Learning and Teaching: policies, procedures and challenges', and the plenary speakers will be Professor Lyle Bachman, Professor Giuliana Grego Bolli, Dr Neil Jones, Dr Waldemar Martyniuk, Dr Michaela Perlmann-Balme and Professor Elana Shohamy. The Call for Papers is now open and will run until the end of January 2011.

For further information about all of ALTE's events and activities, please visit the ALTE website – www.alte.org

To become an Individual Affiliate of ALTE, please download an application form from the ALTE website or contact the Secretariat – info@alte.org

This is free of charge and means you will receive advance information of ALTE events and activities and an invitation to join the ALTE electronic discussion forums.

Studies in Language Testing

October 2010 saw the publication of another title in the *Studies in Language Testing* series, published jointly by Cambridge ESOL and Cambridge University Press. Volume 33, edited by Waldemar Martyniuk, is entitled *Aligning Tests with the CEFR: Reflections on using the Council of Europe's draft Manual*.

In 2003 the Council of Europe released a preliminary version of the *Manual for Relating Language Examinations to the Common European Framework of Reference for Languages (CEFR)*. Over the next five years a wide range of institutions and individuals undertook case studies to pilot this draft version. Towards the close of the piloting phase, a 2-day colloquium was held in Cambridge, UK, enabling practitioners and academics to reflect on and share their experiences of applying the Manual procedures. Insights from this colloquium informed the Manual revision project during 2008/2009.

This volume contains 12 case studies presented at the Cambridge Colloquium in December 2007. They include the linking of a single test to the CEFR, the CEFR-linking of suites of examinations at different levels and large-scale national projects undertaken by examination boards and specialist research institutes from across Europe and further afield. As well as describing their studies and reporting their findings, contributors reflect and comment on their experience of using the draft Manual. A clear and comprehensive introductory chapter explains the development of the CEFR and the draft Manual for linking tests, discussing its relevance for the future.

This volume will be of particular interest to examination boards, language test developers and educational policy-makers, as well as to academic lecturers, researchers and graduate students interested in the principles and practice of aligning tests to the CEFR.

Information on all the volumes published in the SiLT series is available at: www.CambridgeESOL.org/what-we-do/research/silt.html