FROM ONLINE ENTRIES TO ONLINE RESULTS

(DEVELOPING AN INTEGRATED E-ASSESSMENT SYSTEM LINKING INTERNET DELIVERY OF A TEST WITH BACK-END ELECTRONIC PROCESSING SYSTEMS)

Ed Hackett and Paul Seddon

From Online Entries to Online Results

Ed Hackett – Examinations Manger, Assessment and Operations Group; Paul Seddon – CBT Programme Manager, Customer Services Group, University of Cambridge ESOL Examinations (UCLES)

1. Introduction and background

In November 2005, University of Cambridge ESOL Examinations (Cambridge ESOL) launched an internet delivered computer-based version of the Preliminary English Test (PET). Since then, a number of wraparound packages have been introduced to enable centres to make entries and receive results online. In autumn 2007, with the introduction of on screen marking, the final piece of the e-assessment jigsaw will be put in place, providing Cambridge ESOL and its centres with the complete integrated e-assessment package. Further products have now been added to this online delivery system, including tests from other Cambridge Assessment business streams, OCR and CIE (University of Cambridge International Examinations). This paper outlines some of the key development stages undertaken and discusses a number of issues arising out of these developments, both in terms of the questions they raised and the action subsequently taken. It also explores issues that merit further discussion, research or development.

Cambridge ESOL has produced computer-based tests since 2000, but prior to the launch of CB PET in November 2005, these were all CD-ROM based. PET is a general English examination for speakers of other languages and is at level B1 in the Council of Europe framework of reference and Entry Level 3 in the UK National Qualifications Framework. It tests four skills: reading, writing, listening and speaking. Paper-based (PB) PET was introduced in the late 1970s and was most recently updated in format in 2004. With a fast growing candidature, a 45% increase since 2000, and a young exam population, over 70% of candidates aged under 20, it was felt that PET was an appropriate choice of exam for conversion to a computer-based product.

2. Developing and integrated e-assessment system.

With the vast majority of Cambridge Assessment's examinations being paper-based, it was important to development a system which could integrate with existing exams processing systems. This inevitably raises issues with legacy systems. Do you try to enhance the capabilities of the existing system or is it better to bypass it and develop additional software to meet all the necessary

requirements? Often, there is no choice, but to adapt the existing systems, and this can prove both problematic and costly. Furthermore, the issue of IT resource also has to be factored in. Do you wait until there is sufficient resource and budget for every part of the jigsaw to be put into place, or do you develop the product piecemeal, developing the key functional elements first and bringing forward the launch date?

3. Technical Developments

Cambridge ESOL developed its generic online delivery engine, Cambridge Connect, in a phased approach; the primary phase being customer/candidate centric enabling the delivery of a test to candidates over a distributed network. The over-arching requirement was for a delivery engine specifically purposed for the delivery of high-stakes examinations worldwide (i.e. internationally recognised exams with a high surrender value that can used for immigration purposes or school leaving certification for example). As such, there could be no opportunities for a test to be affected by variations in internet connectivity which therefore dictated that whilst the exam could be delivered online, it was downloaded prior to the examination and taken offline.

Cambridge Connect is primarily focussed on test delivery and as such is customer facing; but this is only half the story. Cambridge Connect needs to integrate with back end processing systems such as our Local Item banking System (LIBS) and the Exams Processing System (EPS), which handles candidate entries, marks capture and the processing of results. In addition, Connect integrates with numerous other systems to enable marking and processing from end-to-end in order to create a seamless paperless experience for Centres and Candidates.

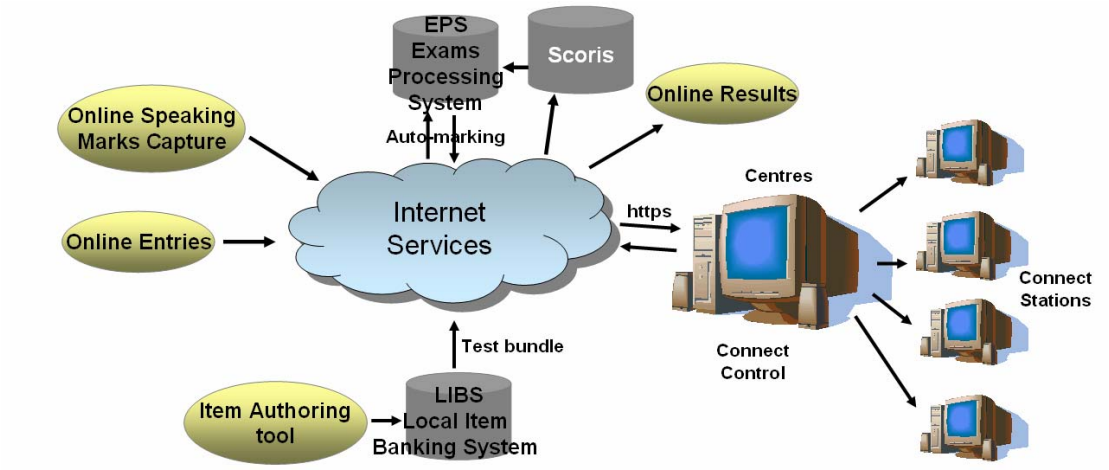

Figure 1. The Connect Framework

3.1 How does it all work?

Within the item bank, pre-tested items are copied into the Item Conversion Tool (ICT). This tools marks up the items in QTi XML, enabling them to be read by the Connect delivery engine, and publishes an electronic test bundle to the Connect hub, a series of web services customers don't see.

For the centres, the experience starts with making Entries, which are keyed in online. Entries are linked to session data in the Exams Processing System and are then communicated along with eligible centre details to the Connect hub.

At the Centre, the Connect software is installed on a network and, at a pre-defined time before the start of the test date, centres can download an encrypted test bundle via https protocols. This test stays encrypted until the test is ready to start on the test day; candidates are provided with login details printed from Connect and start the test. Connect has a number of failsafe features built-in in the event of computer failure. If a candidate's PC fails then the candidate can simply be moved to another PC and resume where they left off. If the Connect Control PC (the PC on which the exam management software runs) fails, a backup recovery tool enables the test administrator to resume the test.

At the end of the test, the candidates' responses are encrypted and uploaded directly to Cambridge web servers at our Data Centre, where different marking applications are employed depending on the type of exam or item types. Some exams consisting of multiple choice question types and short answer responses can be fully automarked; others use the on screen marking application (Scoris), part of Electronic Script Marking system (ESM), enabling examiners to call up candidates' written responses and mark them on screen. Marks are then aggregated and returned into EPS for scaling, grading and results and certificate production.

Cambridge Connect therefore introduces a new and holistic approach (figure 2) to the production, delivery and processing of Cambridge Assessment exams.

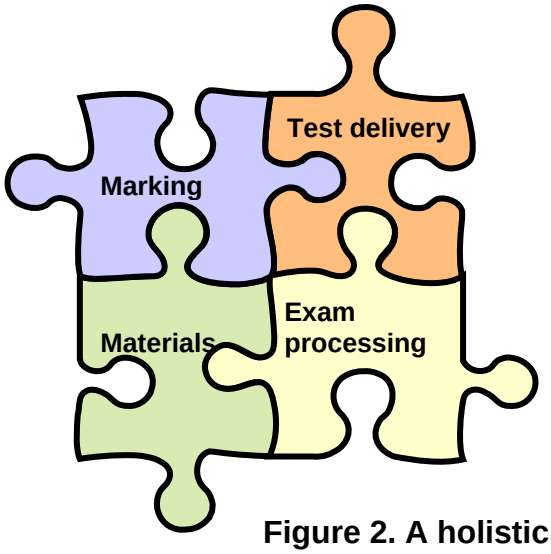

Figure 2. A holistic approach to e-assessment

4. Test Development and Construction

Converting an existing paper-based test for on-screen delivery is very different to the process of developing a computer-based test from scratch. In the latter, you have free rein to develop and trial tasks that you feel best fit this medium. In the former, you have to decide whether the computer-based test is going to follow the same format as the paper-based version and to what extent both modes will be comparable in terms of reliability and results. It was important to Cambridge ESOL that the computer-based test variant of the exam did not advantage or disadvantage candidates when compared to the PB format, and that a grade obtained via the CB mode would have the same value as the equivalent grade obtained using the traditional PB method. A decision was also made to retain the face-to-face format of the Speaking test, though the introduction of online marks capture would allow examiner marks to be keyed into a web application and returned electronically. The key aim was then to prove that it would be possible to transfer the format and task types used in the PB Reading, Writing and Listening tests to an on screen variant.

Four key stages of development were identified:

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

The aim of the feasibility study was to look at the suitability of the tasks in the Reading and Writing and Listening components for on-screen adaptation and to propose designs for trialling. Cambridge ESOL has produced computer-based tests in CD-ROM format since 2000, for example CB BULATS (Business Language Testing Service) and QPT (the Quick Placement Test, which is marketed by Oxford University Press), and development work had already been done on CB IELTS (International English Language Testing System) -launched in May 2005, so a certain amount of knowledge and expertise had already been gained from the development and use of these products.

One of the key issues in converting paper-based materials for on-screen delivery is the use of the computer screen real estate. For example, in a paper-based test the candidate can view two pages of text at one time, whereas a computer screen can only display part of this text at any one time. In addition to this, part of the screen in a CB test is taken up with navigation buttons. This does not present a problem for discrete tasks, tasks with only one item, which can be displayed on screen in their entirety, e.g. PET Reading Part 1 and PET Listening Part 1 (*see table 1 below*), where the task consists of one or more small graphics, one short question and 3 multiple choice options. However, in grouped-question tasks, decisions had to be made over the display of longer text and question input.

Table 1: CB PET Test Content for Reading, Writing and Listening

READING			
Part	**Task Type and Format**	**Task Focus**	**Marking Method**
1	Three-option Multiple choice discrete. *Five* very short discrete texts: signs and messages, postcards, notes, e-mails, labels etc., plus one example.	Reading real-world notices for main message.	Automarked
2	Matching – grouped task *Five* items in the form of descriptions of people to match to eight short authentic-adapted texts.	Reading multiple texts for specific information and detailed comprehension	Automarked
3	True/False – grouped task *Ten* items with an adapted-authentic long text.	Processing a factual text. Scanning for specific information while disregarding redundant material.	Automarked
4	Four-option multiple choice – grouped task. *Five* items with an adapted-authentic long text.	Reading for detailed comprehension; understanding attitude, opinion and writer purpose. Reading for gist, inference and global meaning.	Automarked
5	Four-option Multiple-choice – grouped task. *Ten* items, plus an integrated example, with an adapted-authentic text drawn from a variety of sources. The text is of a factual or narrative nature.	Understanding of vocabulary and grammar in a short text. Reading for general and detailed meaning, and understanding the lexico-structural patterns in the text.	Automarked
WRITING			
Part	**Task Type and Format**	**Task Focus**	**Marking Method**
1	Sentence transformations. *Five* items, plus an integrated example, that are theme-related. Candidates are given sentences and then asked to complete similar sentences using a different structural pattern so that the sentence still has the same meaning.	Control and understanding of Threshold/PET grammatical structures. Rephrasing and reformulating information.	Automarked

2	Short communicative message. Candidates are prompted to write a short message in the form of a postcard, note, e-mail etc. The prompt takes the form of a rubric or short input text to respond to.	A short piece of writing of 35 - 45 words focusing on communication of specific messages.	On Screen marking
3	A longer piece of continuous writing. Candidates are presented with a choice of two questions, an informal letter or a story. Candidates are primarily assessed on their ability to use and control a range of Threshold-level language. Coherent organisation, spelling and punctuation are also assessed.	Writing about 100 words focusing on control and range of language.	On Screen marking
LISTENING			
Part	**Task Type and Focus**	**Task Format**	**Marking Method**
1	Multiple choice (discrete). Short neutral or informal monologues or dialogues. *Seven* discrete three-option multiple choice items with visuals, plus one example.	Listening to identify key information from short exchanges.	Automarked
2	Multiple choice – grouped task Longer monologue or interview (with one main speaker). *Six* three-option multiple choice items.	Listening to identify specific information and detailed meaning.	Automarked
3	Gap-fill – grouped task Listening to identify, understand and interpret information. Using this information to fill *six* gaps on a form or to complete notes.	Longer monologue of neutral or informal nature.	Onscreen Marking
4	True/false – grouped task Longer informal dialogue. Candidates need to decide whether *six* statements are correct or incorrect.	Listening for detailed meaning, and to identify the attitudes and opinions of the speakers.	Automarked

Decisions over the use of pagination, used in the older CD-ROM format tests, and scrolling, the most common format for websites, had to be made. The colour and size of font and background screen colour were also important factors, as was the format of the graphics. Furthermore, onscreen rendering

of the tasks had to be integrated with items drawn from the current paper-based item bank, which meant converting word-based tasks into XML.

The feasibility study revealed that it should be possible to represent all the paper-based tasks on screen and task, navigation and equivalence trialling revealed few major problems. As anticipated, an overall preference for taking PET on computer was expressed by the majority of candidates taking part in equivalence trialling (190 candidates in 4 different countries). 63% preferred taking the Reading and Writing test on computer, as opposed to 20% preferring the paper-based version. For the Listening test, 83% expressed a preference for the computer version, with only 4% preferring the paper test (Hackett, 2005). Candidates found the proposed functionality for answering both multiple choice and typed answers clear and easy to use. Following task trialling, the additional functionality to remove a multiple choice answer already entered was added. This allows candidates to leave a question unanswered, having already entered an answer, should they want to leave it blank and return to it later. It was also discovered that some candidates at this level had difficulty following a grouped listening task and typing answers at the same time (PET Listening Part 3). Candidates were subsequently allowed to make notes on paper and were given additional time to type these up at the end of the task.

No major problems were identified with reading text on screen, though a number of candidates did express a desire to be able to highlight text. This has been backed up by feedback from some candidates taking the test in live sittings, though no drop in reading scores on the CB mode has been identified. Cambridge Assessment is investigating the technology necessary to add this functionality for a future release of Connect. Further research into the impact of reading on screen versus reading on paper is high on the agenda at Cambridge ESOL. In response to the question, 'Did you find reading on computer easier than reading on paper?', 46% found it easier, whereas only 25% preferred reading on paper. This perhaps reflects an increasing familiarity with on-screen reading, at home, in school or at work. PET, as a level B1 test, has a limited reading load for candidates, with the maximum length of text being 450 words. Higher level exams with longer reading passages will exert greater strain on the reader and might impact on the task. Paek (2005), in reviewing CB and PB versions of tests in the American schools sector, noted that extended reading passages tended to appear more difficult in CB format. This is clearly an area warranting further research and the introduction of new examinations to the Connect delivery system will help provide more data for analysis.

Writing also proved more popular on screen, with 67% showing a preference for typing and only 25% expressing a preference for handwriting. CB PET disables grammar and spell checks in an effort to maintain the conditions of the PB equivalent, though the screen does include a word count. However, if we were to attempt to replicate real-life writing situations, it could be argued that grammar and spell check facilities ought to be included. This would necessitate the introduction of a separate markscheme reflecting the resulting improved standards of accuracy and may cause problems in differentiating

between candidates who are naturally able to use language accurately and those who are able to exploit the correction aids available. For the Writing section, other key issues were the impact of typing on candidate performance, and the affect of type-written script on examiner marking; i.e. do examiners treat typed script more harshly or leniently than handwritten script? A number of studies into this area have been carried out for CB IELTS (Thighe et al, 2001, and Green and Maycock, 2004), but given the different test format and candidature, it was agreed that further validation studies would need to be carried out. The benefits of using new marking procedures and analytical tools made available by the advent of on screen marking are explored further in section 5.

5. Marking and Grading

As mentioned above, development of a fully integrated system was split into various phases, with online test delivery preceding electronic marking of responses. The traditional method for marking Cambridge PB tests is via an optical mark reader (OMR) answer sheet. The candidate lozenges in multiple choice answers and writes any written responses within defined spaces on the answer sheet. On return to Cambridge ESOL, written responses are marked either by general markers e.g. for short responses, or by examiners, for longer composition type answers e.g. the letter or story in PET Writing Part 3. The general marker or examiner lozenges a score on the OMR, which is then scanned into the exams processing system, where multiple choice answers are electronically auto-marked against a pre-populated key and added to general and examiner marks. Speaking marks are entered by the examiner onto an OMR and this is returned to Cambridge for scanning. Item level data can then be extracted by the Validation department ahead of grading.

In phase 1 of the project, candidate responses were overprinted onto OMRs so that written responses could be marked in the same way as PB responses, with the OMRs then being scanned. Speaking marks were collected in the same way as for PB (above). The development of an online portal for entering speaking marks at source, in November 2006, meant that speaking marks could be directly ported to the exams database. The introduction of this facility negated the need to print and despatch speaking OMRs to centres prior to the exam and the need for centres to return these marksheets to Cambridge, speeding up the back-end processing of scores and reducing the entry window by 2 weeks.

The final phase of development is the introduction of on screen marking for human rated tasks. This not only allows the speeding up of the marking process, but offers the opportunity for improvements to the examiner marking system, developing online support for markers and contributing to increased rating reliability. In parallel with this system, multiple choice and some short answers will be directly automarked, without the need to print to OMR and scan. The other short productive items, those deemed too complex to be

automarked, will be delivered onscreen to general markers. Longer texts will be routed to examiners, who will undergo co-ordination and standardisation and mark via their home computers. On screen marking has already been developed and used by both OCR and CIE for marking PB products, where completed exams papers are first scanned. For CB, there are obvious cost savings, as responses do not require scanning. The responses returned via Connect are displayed to the examiner using same screen view that the candidate sees.

However, one of the additional advantages of using on screen marking is not simply savings in time or cost. It is the opportunities it offers for the implementation of new examiner marking models, that is particularly interesting. There are various models employed for examiner marking of Cambridge ESOL exams, utilising both on-site and at-home marking scenarios. PB PET is currently marked on-site using a partial remarking model. Examiners are put into teams which are monitored by a team leader, who in turn reports to a Principal Examiner. Following co-ordination and standardisation, each examiner is monitored by the team leader, who informs the examiner of leniency, harshness or erratic performance early on in the process. The aim of this approach is that performance is monitored and modified where necessary. Batches of scripts are then remarked where appropriate and monitoring continues over the marking weekend. At home marking models also include co-ordination and standardisation, in addition to batch sampling. Examiner marks are then subject to scaling, to take account of identifiable leniency or harshness. A third model is double marking, with both examiner marks being averaged, or those deemed outside acceptable tolerance, i.e. differing by too great a margin, being sent to a third, experienced, rater.

On screen marking offers the opportunity for a new model of marking and the possibility of greater intervention in examiner marking behaviour. In addition to the use of co-ordination and standardisation scripts as processes designed to appropriately align examiner behaviour, there is also the possibility of using seeded ‘gold standard’ scripts (Shaw, 2007) as a means of monitoring such behaviour. Gold standard scripts are candidate samples specially selected as models for use in blind monitoring. These scripts are selected and pre-marked by the PE and a group of senior team leaders, and then seeded as ordinary unmarked scripts into the marking pool each examiner gets. The Principal Examiner or Team Leader is then able to monitor, at various stages during the marking, the relationship between the agreed marks for these scripts and those given by different examiners, and feed this information back into the marking process as a means to achieving greater reliability between markers. Furthermore, the electronic capture of interim as well as final marks provides the validation group with valuable information that can feed into future research. Shaw (2007) identifies a number of interesting research questions that would benefit from the capture of this data:

- In what ways do raters differ? Is there a gender effect? (Facets of Rater Status, Rater Profile and Rater Behaviour.)

- Is it possible to identify distinct rater types and certain patterns of rater behaviour? (Facet of Rating behaviour.)
- What amount of training/re-training is required? Can training improve raters' self-consistency? (Facets of Rater Behaviour and Rater Training.)
- How does assessment differ when marking electronically as opposed to paper-based marking? (Facets of Rater Profile and Rater Behaviour.)

Shaw goes on to state that the data gathered from such exercises could also be used to establish whether particular raters favoured candidates from a particular L1 background or could be used to investigate further the relationship between the tasks, the candidates and examiners. In PET Writing Part 3, candidates are given a choice between writing a letter or a story. We can now investigate further the question of whether rater reliability varies according to the task, and if certain examiners have greater reliability marking one task type as opposed to another. It may then be possible to allocate certain task types to particular types of raters.

On screen marking for CB products using the Connect delivery engine is scheduled for autumn 2007, so we are unable to comment on the live implementation of this software. On screen marking will, however, provide the final link in our online delivery and processing system, leading to a fully integrated e-assessment package.

6. Future development and research

Computer-based assessment using a system like Cambridge Connect raises a multitude of research opportunities that are likely to impact on the way we assess candidates in the future.

With computer-based assessment we have a clear insight into the examination process from a candidates' point of view that until now has been impenetrable. We can log each and every key stroke a candidate makes and are able to determine:

- which questions a candidate attempted first
- which questions a candidate returned to, changed their answers etc
- how long a candidate spent on each question
- whether two candidates sitting next to each other input the same answers at the same time

6.1 Where might this take us?

Cambridge Connect, together with the on screen marking application, will provide a wealth of information for formative assessment, for building diagnostic assessments and providing scaffolding to help the candidate. It could enable assessment organisations to measure candidates' abilities not

just by getting the answer correct, but also on how long it took the candidate to come up with the correct answer and therefore award additional marks for speediness. It enables assessment organisations to pinpoint a candidate's ability or knowledge by tracking which tasks candidates struggled with or conversely, which tasks are not measuring or performing well in a test because a whole cohort struggled with it. Furthermore, the possibilities for live item calibration (live pre-testing) by seeding uncalibrated tasks into a live exam offered by computer-based assessment enables both exam boards and candidates to reap the benefits and achieve even more meaningful measurement of candidates and their abilities.

7. Conclusion

In developing Cambridge Connect and integrating it with both existing processing systems and newly developed wraparound e-services, Cambridge Assessment can now deliver high stakes examinations worldwide, achieving vastly reduced entry and results processing times. We are also in the position to explore more fully the comparability of computer-based tests with their traditional paper-based equivalents, and how the differing modes impact on both candidate and examiner behaviour and performance. As Jones (2007) states, 'It is important for Cambridge ESOL to define an approach to comparability which will guide the validation of …(new CB examinations using Cambridge Connect), … while providing a more general framework for thinking about comparability of technology-based and traditional assessment.' It therefore hoped that a greater understanding of the candidate experience, in terms of their interaction with computer-based tests, will inform the development of future computer-based tasks and tests.

8. References

Green, A and Maycock, L (2004) *Computer-based IELTS and paper-based IELTS, Research Notes 18*, November 2004 (UCLES).

Hackett, E (2005) *The Development of a Computer-based version of PET*, Research Notes 22, November 2005 (UCLES).

Jones, N (2007) *The comparability of computer-based and paper-based tests: goals, approaches, and a review of research*, Research Notes 27, February 2007 (UCLES)

Paek, P (2005) *Recent Trends in Comparability Studies*, PEM Research Report 05-05, Pearsonedmeasurement.com/research/research.htm

Seddon, P (2005), *An overview of Computer-based PET*, Research Notes 22, November 2005 (UCLES)

Shaw, S.D (2007) *Modelling facets of the assessment of Writing within an ESM environment*, Research Notes 27, February 2007 (UCLES).