From Paper to Screen: some issues on the way

Paper presented at the International Association of Educational Assessment conference, 13th – 18th June 2004.

Nicholas Raikes, Jackie Greatorex and Stuart Shaw

University of Cambridge Local Examinations Syndicate (UCLES[1]), 1 Hills Road, Cambridge CB1 2EU, United Kingdom.

Email:
n.raikes@ucles-red.cam.ac.uk | greatorex.j@ucles.org.uk | shaw.s@ucles.org.uk

Abstract

In the UK and elsewhere some examination agencies have programmes to move from an entirely paper-based examination system to a hybrid where paper scripts are scanned and digital images are distributed electronically for on-screen marking (scoring). This hybrid is seen as a way of realising some of the benefits of digital scripts in a context where paper is likely to remain important for many years to come.

The University of Cambridge Local Examinations Syndicate (UCLES) first tried on-screen marking of scanned paper scripts in 1999. Results from this and subsequent trials were encouraging but inconclusive, and recently UCLES and its UK-trading subsidiary OCR launched a comprehensive programme of research, development and evaluation. In the present paper we share some of the issues that we believe will need to be investigated during the programme, and present results from UCLES' first two studies.

Acknowledgement

Many people from UCLES, from RM plc – our lead technical partner for on-screen marking – and from ETS, whose Online Scoring Network systems we used, contributed to the work reported here.

[1] The UCLES Group provides assessment services worldwide through three main business units.

- Cambridge-ESOL (English for speakers of other languages) provides examinations in English as a foreign language and qualifications for language teachers throughout the world.
- CIE (University of Cambridge International Examinations) provides international school examinations and international vocational awards.
- OCR (Oxford, Cambridge and RSA Examinations) provides general and vocational qualifications to schools, colleges, employers, and training providers in the UK.

For more information please visit http://www.ucles.org.uk

Introduction

Computer assisted assessment offers many benefits over traditional paper methods, but paper has traditionally been such an integral part of many instructional programmes – and of the high stakes, mass assessment systems that support them – that any such assessment system – particularly one that makes as heavy use of constructed response questions as UCLES does – will have to deal with paper for many years to come.

In consequence some assessment agencies have in place programmes to support the digitisation, electronic distribution and on-screen marking[2] of paper scripts.

UCLES is investing in on-screen marking of scanned paper scripts as part of a much wider strategy to re-conceptualise what we do in the digital age. We recognise the continuing need to handle paper, albeit as a perhaps diminishing part of an integrated system handling a variety of assessment evidence collected using the most appropriate media and processes.

UCLES held its first trial of on-screen marking of scanned paper scripts (hereafter referred to simply as on-screen marking) in 1999, and in 2000 and 2001 we conducted a series of major investigations involving international A level, O level, IGCSE and English as a Second or Other language examinations. We found evidence to suggest that examiners' on-screen marking of short answer scripts was reliable and comparable to their marking of the paper originals, and that the images were legible (in fact in some cases examiners reported that scanned scripts were more legible than the originals, since faint writing was darkened by scanning and examiners could magnify small writing). We concluded, however, that more research was needed, particularly concerning extended responses, to ascertain in exactly what circumstances on-screen marking was valid and reliable.

In 2003 UCLES returned to on-screen marking with renewed interest and subsequently partnered with RM plc. Interest was renewed perhaps as part of a realisation that paper would be a feature of our market for some time to come and that we could not wait for examinations to go "on-line" before providing our customers with some of the enhanced services that digital processing could support. UCLES remains determined, however, to involve stakeholders every step of the way, and to research comprehensively the impact of any proposed changes so that any effects on reliability or validity are fully understood.

UCLES has launched a major research programme to support its introduction of on-screen marking, and in the present paper we give details of some of our current thinking and plans, although these are still very much under development. In January 2004 our first study investigated the effects on schools and colleges of two different ways of providing machine readable script identifiers. We present a brief summary of the findings beginning on page 15 of the present paper. In March 2004 some senior examiners took part in our first, exploratory on-screen marking session of the current

[2] There are many different terms used for 'marking'. Some examples are 'scoring', 'reading' and 'rating'. There are also many terms for 'on-screen marking'. Some examples are 'e-marking', 'on-line marking', 'on-line scoring' and 'on-screen rating'. The situation is similar for the term 'examiners' who are referred to as 'readers', 'raters', 'judges', 'markers' and so on. When references are made to other authors' work, their terms are used. Elsewhere the terms 'marking', 'on-screen marking', 'markers' and 'examiners' will be used.

programme. We used ETS's comprehensively researched and tested Online Scoring Network (OSN) software and systems without any modifications, the aim being to involve examiners at the earliest possible stage of the current programme so that they might shape any modifications required for our examinations and help steer how on-screen marking should be introduced. The study produced a large amount of very valuable data, and we provide a very brief summary beginning on page 17 of the present paper. Our next trial will involve a much wider range and number of examiners and will use scripts scanned in June 2004. Planning for this trial is still underway, though a revised version of the OSN software will be used that takes into account findings from the previous trial.

We will continue to publish our research plans and findings as they develop, and welcome feedback and suggestions. As research findings and business plans build and combine we will gain a clearer picture of the ways in which we may introduce on-screen marking. If the benefits of on-screen marking are confirmed and the risks manageable then UCLES plans to introduce limited on-screen marking in November 2004. All going well, wide scale on-screen marking may be taking place in 2006.

Plan for the present paper

In the present paper we focus principally on assessment issues and generally ignore issues relating to technology and cost, which are being investigated by other colleagues.

The structure for the rest of the present paper is as follows.

After a literature review we summarise UCLES' existing, paper based systems. We then consider changes that on-screen marking might support, and the issues for research that these changes would present. After a conclusion we provide brief summaries of the two initial studies referred to above.

Literature review

Zhang et al (2003) reviewed the literature. They cited Powers et al (1997), Powers and Farnum (1997) and Powers et al (1998), whose work is relevant to the present paper. A very brief overview of this work, as described in Zhang et al (2003), will now be given. In a pilot study by Powers et al (1997), experienced readers scored essay responses on paper and using ETS' On-line Scoring Network (OSN). The readers who used OSN were fairly positive about on-line scoring. There were no differences between the average scores awarded in either medium and inter-reader agreement was comparable for paper and on-line scoring. Powers and Farnum (1997) gained similar results – they found that the medium in which essays were presented to readers, on-screen or paper, did not affect scores. For experimental purposes Powers et al (1998) relaxed some of the academic credentials traditionally required of readers. They found that after training, a good proportion of inexperienced readers exhibited an equivalent level of accuracy to that of experienced readers. This indicated that the prerequisites for readers could potentially be relaxed without sacrificing the accuracy of scoring.

In England, Newton et al (2001) evaluated an on-line marking trial of "year 7 progress tests". These externally marked tests were introduced in 2001 to monitor the progress in year 7 of 12 year olds who had failed the previous year to reach the level expected of the majority of 11 year olds in England's National Curriculum tests in Mathematics and English. The scripts were scanned for the trial at the item level and the external markers marked them using NCS Pearson software in a central marking venue. The images were presented to the markers using the Internet and supervisors were available to help the markers. The items were divided into different item types: 'data entry', requiring unskilled markers; 'clerical', requiring semi skilled markers and 'expert', requiring skilled markers. The data entry and clerical items were double marked and discrepancies resolved by a senior marker. A sample of the experts' marking was double marked. The authors concluded that the marking centre based model had potential if examiners embraced the culture change (currently they mark at home). Although the on-line marking in their research took place in marking centres they acknowledged that a major advantage of a web based on-line marking system was that examiners may mark at home. They recognised that there were hardware and software obstacles to implementing such a system and suggested that the loss of face to face interaction might affect quality.

Whetton and Newton (2002) reported on the same trial. There were high correlations between on-line marking and conventional marking for all marker types and examinations except for 'writing' and 'spelling/handwriting', both marked by experts on screen and on paper. These lower results were not necessarily caused by the new technology; it could be that the markers were not as expert as had been hoped. The conventional marks given by all marker types were on average a little higher than the on-line marks. There is no definitive explanation for this effect, which does not accord with previous research findings, but it is clearly important, and warrants further investigation. The differences meant that 5% of maths and 27% of English candidates would have received a lower level, had they been marked on-line. It was found that using non-expert markers for non-expert questions was technically effective, a similar finding to that of Powers et al (1998), noted above.

Whetton and Newton (2002) also found that there were a high number of candidates' responses outside of the image presented to markers. This was a characteristic of particular pupils and if this approach of presenting images to markers were continued these candidates might be disadvantaged. However, they found that for their data there was only one paper – mathematics – where answers beyond the images might have effected the discrepancy between conventional and on-line marking. Examiners were not happy marking a response to just one item as they thought the candidate might have written more in response to the item than they actually saw. The pilot illustrated that large numbers of scripts could be scanned and that the clerical collation of marks was rapid and accurate. However, "Marking is not simply an administrative process, it involves issues relating to the validity of the assessments and also judgements by human beings. The computerised features of the system must serve in support of providing the most accurate realisation of the candidates achievement, not only provide speed and cost reductions." (Whetton and Newton, 2002, 33).

Sturman and Kispal (2003) undertook research following on from the work of Whetton and Newton (2002) and Newton et al (2001). Sturman and Kispal (2003) compared electronic marking and paper-based marking with the aim of establishing whether e-marking could be a viable means of gathering the data required of a pre-test. Their work was in the context of marking pilot items in tests of reading, writing and spelling for pupils typically aged 7 to 10 years. Their analysis explored marking effects at the test, item and pupil levels. An analysis of mean scores showed no consistent trend in scripts receiving lower or higher scores in the e-marking or paper marking. The authors point out that neither the e-marking or paper marking scores can be considered to be the true score. They add that the “absence of a trend suggests simply that different issues of marker judgement arise in particular aspects of e-marking and conventional marking, but that this will not advantage or disadvantage pupils in a consistent way” (Sturman and Kispal, 2003, 17). They also found that e-marking is at least as accurate as conventional marking. When there were discrepancies between e-marking and paper-based marking these generally occurred when the marker judgement demands were high.

Sturman and Kispal (2003) noted that when marking on paper a pupil’s performance on the test as whole may influence a marker’s judgements about individual items. This cannot occur when marking individual items on screen so e-marking is arguably more objective. They suggested that there should be more research regarding the comparability of data from paper and e-marking at the pupil level.

Zhang et al (2003) compared scoring on-line versus scoring on paper using student responses from the Advanced Placement (AP) program, and ETS’s Online Scoring Network (OSN) system. AP examinations include multiple choice questions and an essay section. In AP tests the candidates complete paper answer booklets by handwriting their essays and responding to multiple choice questions by shading in boxes. This study is particularly relevant to UCLES – whose examiners currently mark at home – since it compared paper-based marking in a central location with OSN on-line marking in remote locations. At the item level there were statistically significant differences between the mean scores from each scoring environment but the differences were equally likely to favour remote scoring or paper and pencil scoring. The agreement between OSN readers on free response questions was "at least as good as that for those who read in a traditional operational setting" (Zhang et al, 2003, 21). There was no statistically significant difference between the two scoring environments in terms of internal reliability or inter-reader agreement. Zhang et al, (2003, 21) concluded that "the results obtained from OSN are extremely similar to those obtained with traditional AP scoring methods."

Zhang et al (2003) also surveyed the readers who had taken part in the study. They found that the readers generally rated *OSN specific* training as being effective in training them to use OSN, and *subject specific scoring* training as effective in helping them to score accurately. Readers who gave OSN training a negative rating drew attention to, for example, the lack of discussion with other readers, there being no option to print commentaries to training essays and having to scroll to read the essays. Most readers sought technical help which they thought was successful. Readers' reactions to most aspects of OSN were at least satisfactory, but a "significant minority" of readers rated the handwriting image display to be less than satisfactory. Generally readers consulted their scoring leader at least once and telephones were rated as satisfactory for this communication. Nearly half the respondents to the questionnaire reported difficulty connecting to the OSN website and 38% had trouble with slow download speed. Scoring leaders generally thought the OSN training was effective and that the telephone was at least satisfactory for discussing scoring issues. They too drew attention to the inability of readers to interact with each other. 75% of respondents who were scoring leaders encountered trouble connecting to the OSN website and 50% reported problems with download speed.

In a paper considering numerous aspects of comparability between computer assisted and conventional assessment, Bennett (2003) considered whether scoring presentation (e.g. on paper or on-screen) affects the scores given by raters. After reviewing the literature he concluded that "the available research suggests little, if any, effect for computer versus paper display" (Bennett, 2003, 15). When he came to this conclusion his review did not include Whetton and Newton's (2002) findings of consistently higher marks being credited when marking on-line as opposed to on paper.

Twing et al (2003) compared the marking of paper and electronic images of essays. The allocation of markers to groups was controlled to be equivalent across the experimental conditions of paper and electronic marking. The authors concluded that the statistical evidence indicated that the paper based system was slightly more reliable than the image based marking. They surveyed markers and found that some had never "interacted" with a computer before and that there was some anxiety about marking on-screen. They also found that image based markers finished faster than paper based markers.

In summary, the literature – though far from comprehensive – suggests that on-screen marking may prove to be as reliable and valid as paper based marking. This finding may well depend however on the way in which on-screen marking is implemented and on the context. There appears to be some scope to replace expert markers with less qualified personnel in some circumstances without compromising marking quality. Research into examiners' experiences of and views about on-screen marking had mixed findings, and these should be explored fully as part of the development of any new system.

Our current system

Different Business Streams within UCLES operate in different ways, but the following features are typical of many examinations run by OCR and CIE. Cambridge-ESOL's procedures are markedly different in several areas and will not be considered here.

Candidates enter through examination centres which are generally the schools and colleges where they are taught. Dedicated test centres are not widely used.

Although in some cases CIE examines every eligible person within a country, generally entries are not representative of any population and each examination session there is usually some "churn" in the centres entering candidates.

Most items require a constructed response with lengths varying from one or two words up to extended writing and encompassing graph sketching, diagrams, etc.

Candidates either answer on the question paper or, typically for papers involving long answers or question choice, on separate answer booklets. Additional sheets and graph paper may be attached.

Question papers have identifiers printed on them that indicate the examination, but candidates must write on their names, candidate number and centre number. Answer booklets and additional sheets contain no printed identifiers and candidates must write in an examination identifier as well as the other information mentioned previously.

Scripts are generally sent to Assistant Examiners who mark (score) them at home. Each Assistant is part of a team headed by a Team Leader, who reports in turn to the paper's Principal Examiner, who reports to the Chief Examiner of all the assessment components of a syllabus. Each examiner will have been pre-allocated scripts from one or more centres based on the entries and scripts will either be sent to them directly by the centre (OCR) or via UCLES (CIE). Examiners are typically qualified and experienced teachers of the subject they are marking.

All examiners are required to attend a co-ordination or standardisation meeting at the start of the marking period where mark schemes are finalised and examiners' application of them standardised. All examiners mark and discuss a sample of photocopied scripts at the meeting. Examiners also raise and discuss any un-anticipated responses that they have observed in their own script allocations. Mark schemes are modified if necessary and finalised. After the meeting, Assistant Examiners may begin provisional marking but must send a sample of ten marked scripts to their Team Leader for re-marking. The Team Leader may approve the Assistant or request a further sample. When approved, an Assistant must go back over any provisional marking and make any changes necessary in the light of the Team Leader's guidance and proceed with new marking. No further feedback is given to Assistants during marking, though at least one further sample of scripts chosen from each Assistant will be re-marked.

When a script has been marked, the examiner adds up the marks and transcribes the total only onto a machine readable form for scanning into UCLES' systems. UCLES does not usually keep records of marks at less than whole paper level.

When Assistant Examiners have finished their marking, senior examiners and UCLES officials meet to consider whether any Assistant's work needs to be re-marked (if the Assistant was erratic) or scaled (if the Assistant was systematically too lenient or severe). Evidence considered includes Team Leaders' re-marking of sample scripts and recommendations, statistical evidence and additional samples of Assistants' work.

Since most of the items and examination papers used by CIE and OCR are not pre-tested and calibrated, grade cut scores for each examination are not set until after marking is over. The process of setting cut scores is called Awarding, and CIE and OCR follow slightly different procedures.

CIE's Awarding process is as follows. After marking, the paper's Principal Examiner, sometime after consultation with Team Leaders and Assistant Examiners, recommends cut scores based on his or her view of the standard of candidates' work and with reference to the standards set in previous years and grade descriptors. These recommendations are considered alongside statistical and other evidence (e.g. comments about the paper from teachers) at an Awarding Meeting. Statistical evidence usually includes the mean, standard deviation, and mark distributions for this and other papers taken by the same candidates in the syllabus, the grades centres estimated for the candidates, and similar data from previous years. Cut scores recommended by the Awarding Meeting are finalised by CIE's Standards and Projects Manager.

OCR's Awarding process differs mainly in the weight given to looking at sample scripts in the Awarding Meeting. First, the Principal Examiner recommends (after marking and possibly after consultation with Team Leaders and Assistant Examiners) a range of marks in which the cut scores are expected to lie. An Awarding meeting is held, usually attended by the Principal Examiners of all papers within a syllabus, the Chief Examiner responsible for the syllabus as a whole, and OCR officials. The meeting looks at sample scripts within the recommended mark ranges, archive scripts exemplifying standards set in previous years, and statistical and other evidence from this and previous years, often including grade descriptors and teachers' comments.

The meeting recommends grade cut scores to a subsequent Grade Endorsement Meeting for final checking.

After Awarding, there is a final process of checking known as Grade Review, where scripts are re-marked by senior examiners based on their proximity to key grade cut scores and evidence of possibly anomalous marking. Grade Review occurs at a central location near where returned scripts are stored.

Possible changes in an on-screen environment

On-screen marking may require or afford numerous changes to current procedures. Some of these are briefly described below.

Question Papers and Answer Booklets

When scripts are scanned the system must be able to identify who wrote a script, which centre they entered through and which question paper they were attempting. When candidates answer on the question paper it is easy to pre-print a bar code identifying the question paper, but larger changes are needed to cater for examinations involving answer booklets and to provide machine readable centre and candidate identifiers. Early on in the on-screen marking programme UCLES decided to try two different methods of recording machine readable script identifiers. In one way – Fully Pre-personalised (FP) – all identification details would be pre-printed onto scripts as bar codes, in the other way centre and candidate identifiers (and a question paper identifier for answer booklets) would be read directly from candidates' handwriting using Intelligent Character Recognition (ICR). UCLES conducted research comparing the FP and ICR approaches in January 2004, and a brief summary of this research is included in the present paper, beginning on page 15.

Before scripts may be scanned the pages must be separated by guillotining the central seam to remove staples. Anything written by candidates in the extreme centre of a double page will be lost, and UCLES will print a dark band down this central area to prevent candidates from writing there.

Once a script has been scanned, the system must be able to identify which part(s) of which image(s) contain the answer to a question. This is so that the correct mark entry boxes may be displayed with the answer when it is marked, and in case different questions are to be sent to different examiners for marking. Question identification is relatively easy to do when candidates answer in defined places on a question paper, and the papers may only require minor changes – or even no changes – to encourage candidates to write in an appropriate place. Answer booklets present more of a challenge, and more substantial changes may be required to encourage candidates to clearly label and separate their answers.

One reason for splitting answers might be so that they may be distributed to markers according to the expertise needed to mark them. Some questions – and the corresponding marking guides used by examiners – might be susceptible to being modified so that they could be clerically marked, though assessment validity requirements remain paramount.

Who marks what?

One of the key potential benefits of on-screen marking is the flexibility offered in terms of distributing candidates' work for marking. Scripts might be split by question and the various questions distributed as appropriate to:

- trained and standardised clerical markers, i.e. markers who have little or no subject knowledge;
- trained and standardised markers with subject knowledge, for example recent graduates or postgraduate students in a suitable subject;
- trained and standardised Assistant Examiners.

Moreover, a marker may be sent the work of any candidate, and there is no logistical requirement for centres' scripts to be distributed together.

Training and co-ordinating markers and assuring marking quality

When scripts are scanned, the images may be copied and distributed at will. Many (or all) markers may be sent copies of the same answers for training and co-ordination without recourse to photocopying. Additional training and co-ordination answers may be sent whenever necessary, and markers may almost instantly pass a difficult or non-standard answer to senior colleagues. This may be particularly useful in the early stages of marking when mark schemes are still being finalised. Face to face co-ordination meetings involving all examiners might even prove unnecessary if on-screen marking is coupled with digital communication tools and, possibly, local team meetings.

Quality assurance procedures may be revolutionised by on-screen marking. For example, Team Leaders or other senior examiners may instantly call up samples of a marker's work, and may easily direct some scripts or answers for double marking. On-screen marking also affords new opportunities for "background" monitoring, allowing senior examiners to target their quality assurance work more effectively. For example, during the marking period some answers or scripts may be sent to everyone marking the relevant questions and the item-level marks compared. Frequent differences beyond an acceptable tolerance may indicate individual examiners who need more guidance (if most examiners agree on the marks) or mark scheme deficiencies (if there is lots of disagreement). Either way some previously marked answers may need to be re-marked when the problem is corrected, and this is easy to arrange with on-screen marking. So called "gold standard" answers or scripts, where marks have been pre-agreed by senior examiners but which appear as normal to markers, may also be included throughout the marking period. These might be used in particular for background monitoring of less experienced markers – a marker's Team Leader could be warned if differences between the marker's marks and the gold standard marks exceed a certain threshold, and the Team Leader may inspect the marker's work.

On-screen marking also enables more useful statistical information to be provided to those responsible for assuring marking quality. For example, the practical requirement to keep all the paper scripts from one centre together need no longer constrain the number of centres from which a marker's work comes. Images may be drawn almost at random, and consequently differences between the mean marks awarded by markers marking the same items (but different candidates) may be tested for statistical significance. Since item marks are collected this could be done for each item as well as for aggregated marks. If scripts are split so that different markers mark different bits then this also provides useful data for comparing examiners, since a strong correlation is expected between candidates' performance on different bits of a script. With some overlap between markers, either provided by sharing a candidate's work between more than one marker or through some double marking, the item level marks may be analysed according to the Rasch model or Item Response Theory to place all marker severities / leniencies onto a common scale regardless of the candidates or items they marked. Analysis of the residuals from this modelling may prove even more valuable since misfit may be tested for statistical significance and may indicate aberrant marking. Of course statistical analysis cannot prove marking irregularities, but the results may be used to guide supervisors to scripts or markers that warrant review.

Perhaps the key quality assurance benefit of on-screen marking is the ease with which possible marking problems may be detected and investigated early and interventions made quickly. If severe or lenient marking may be eradicated, marker scaling would no longer be necessary. Even if this proves difficult, scaling may be investigated and applied at the item level if appropriate. With quicker and easier detection of aberrant marking and the ability to have script images sent for re-marking instantly, Grade Review marking after Awarding may be reduced or eliminated, giving more time for these checks. Even if some post Awarding re-marking still proves necessary, senior examiners need not do it where the scripts are stored but may work at home, reducing costs and inconvenience.

Awarding

Those responsible for setting grade cut scores may be helped by the new statistical information and improved access to scripts that the new digital environment may provide. With item level marks available Awarders may see, for example, how average marks on any item vary between candidates with different paper totals (or indeed between groups of candidates selected according to any available criteria). In this way Awarders may focus on the key discriminating items for a particular grade, and may easily view sample answers drawn from scripts with, for example, a particular total mark. Of course they may also want to make holistic judgements about whole scripts, and in a digital environment easy access to scripts drawn according to any available criteria may be facilitated.

Item level information may help Awarders identify items that did not perform well, and if desirable it may be possible to exclude these items from candidates' scores.

The Awarding process may also be changed if scripts are split up and different bits marked by different markers. If nobody marks whole scripts, Principal Examiners may have to change the basis on which they recommend grade cut scores. Indeed the whole way in which judgements about different pieces of evidence are combined during Awarding may change in a digital environment.

Reporting and feedback

The collection of item level marks will enable richer feedback to centres and candidates. Centres may be provided with information about how well their candidates did on each question or topic area, compared with other candidates, and this may help centres identify the strengths and weaknesses of their teaching. Similar information could be provided for individual candidates, if there is a demand for it.

Item statistics may prove to be extremely useful feedback to paper setters (the people who write question papers). Setters may identify questions that did not perform as expected, and consideration of the questions and candidates' answers to them may help them improve future questions.

Issues for research

Centres and candidates

What are the effects on centres of different ways of providing machine readable script identifiers? Depending on the method adopted, centres may have to change some procedures. For example, if fully pre-personalised stationary is used centres must ensure that each candidate receives the correct stationary and must also store extra pre-personalised stationary for each candidate in case it is needed. Which method do centres prefer? Our first piece of research in the current programme involved us working with several centres during the January 2004 examination session to investigate the impact on them of two alternative script identification methods – fully pre-personalised and ICR, described above – that UCLES was considering. A brief summary of this research is included in the present paper, beginning on page 15.

What additional feedback, derived from item level marks, do centres and candidates value, and how should it be presented? How best should we provide them with online access to scripts?

Examiners

Examiners are central to UCLES' work, and they must be fully involved and consulted.

How do we identify and respond to the needs, concerns and aspirations of Assistant Examiners, Team Leaders, Principal Examiners and Chief Examiners?

What training will examiners need? How will examiners' access to computer equipment and Internet connections be provided?

How should examiners' fees and expenses be changed to be fair in the new environment?

How do we support examiners so that they themselves may take a leading role in shaping the new environment?

How do we retain existing examiners and recruit new ones?

Question papers and mark schemes

What question paper and answer booklet designs are most effective at encouraging candidates to write correctly labelled answers in appropriate places, using appropriate materials?

What changes to question papers and mark schemes may be made to facilitate marking by clerical or graduate markers or, indeed, by automatic methods?

What constraints and opportunities relating to item design are associated with on-screen marking?

How are constraints and opportunities best communicated to the setters who write the papers?

What feedback from previous papers is most useful to setters, and how is it most effectively presented to them?

Can all existing paper-based examinations be marked on screen or are there features of question papers which cannot be accommodated or which are too costly to accommodate?

What changes are sensible given a possible future migration to computer based testing (i.e. where candidates answer using a computer)?

What changes are acceptable to setters and other stakeholders? What changes do they desire?

What are the effects of changes on validity and reliability?

Marker Training and Co-ordination

What training and co-ordination methods are appropriate for clerical markers, graduate markers and examiners?

How should training and co-ordination methods vary by subject, level and item type?

How may training and co-ordination best be done in an on-screen environment supported by digital communication tools? In what circumstances are face to face meetings appropriate? Should computers be used at face to face meetings, and if so how, and what are the practical constraints?

If examiners are co-ordinated using paper scripts, does this transfer to on-screen marking?

How best may a community of practice amongst examiners be supported in a digital world?

If better communication between examiners is fostered, will aberrant examiners negatively influence the marking quality of other examiners?

How will the training and co-ordination roles of senior examiners, Team Leaders and Assistant Examiners change?

How should evidence about candidates' answers be collected and communicated so that mark schemes may be finalised? If some examiners are no longer involved in this, will they feel marginalised?

What changes are acceptable to examiners and other stakeholders? What changes do they desire?

What are the effects of changes on validity and reliability?

Marking

When should clerical markers, graduate markers, or examiners be used?

How much of a script should each marker see? How does this vary by marker type, item type, subject and level?

When marking is investigated, the findings may be influenced by, for example:

- Item types, subject, level and examining personnel;
- The marking software and the choice of computer equipment and Internet connection.
- Marker training, co-ordination, quality assurance and feedback to markers.
- Whether a marker marks whole or partial scripts.
- Whether clerical, graduate or Assistant Examiners are used, and in what proportion.

The above factors must be borne in mind – and controlled – when designing research to answer the following questions.

Are marks produced at the end of the process acceptable in terms of validity and reliability? Are they as good – or better – than those produced through other (including conventional) processes?

Are there any systematic changes to severity or leniency? If so, may they be corrected for post hoc?

What are markers thinking about when they are marking? Does the marking application, computer equipment or Internet interfere with markers' cognitive processes? What are the effects of scrolling and different screen resolutions?

What annotation facilities should be provided? What are the effects of different levels of annotation on: (a) a marker's marking process, (b) marking supervisors' (e.g. Team Leaders) ability to monitor and advise a marker, and (c) validity and reliability?

How often is part of a candidate's answer rendered hard to read or lost as a result of the scanning or clipping processes or misdirected as part of an item separation process? (NB: this will be influenced by changes to question papers and answer booklets). What are the effects in terms of marks? What are the effects on markers' thoughts (and therefore marks) of clearly missing work? How often do markers *suspect* that work has gone missing and what effect does this have on their marking?

How long may markers work without a break before marking quality deteriorates? (NB: Health and safety issues must also be addressed). How productive are markers?

Does marking location (i.e. whether at home or in a marking centre) affect validity, reliability or productivity?

Quality assurance

How should quality assurance procedures vary by item type, subject, level and type of marker?

Should markers be monitored more tightly in the first few days of live marking compared with later periods, and how should monitoring intensity change over time? Should monitoring intensity depend upon a marker's experience and previous performance?

What is the most effective combination of double marking (where neither marker sees the other's marks or annotations), re-marking (where the second marker does see the first's marks and annotations) and multiple marking (either using "gold standard" answers that have been pre-marked by senior examiners, or using previously un-marked answers distributed to many markers)?

What criteria should be used when selecting answers for gold standard or multiple marking use?

How will the quality assurance roles of senior examiners, Team Leaders and Assistant Examiners change? What additional roles are created?

What statistical information should be calculated? Who needs what, and how and when should it be communicated?

What software facilities are required?

What are Team Leaders and other quality assurance personnel thinking about when evaluating a marker's work? Does the software, computer equipment or Internet interfere with their cognitive processes?

What feedback should be provided to markers? When and how should it be communicated? What balance should be struck between automatically generated feedback, human feedback, group level feedback and individual feedback?

Is there evidence that feedback during marking may lead to undesirable disturbances in markers' behaviour?

What are the cost and productivity implications of different types and levels of quality assurance?

Is marker scaling still required? If so, what are the implications of introducing item level marker scaling?

Is grade review marking still required? If so, under what circumstances is it necessary?

What changes to quality assurance practice are acceptable to examiners and other stakeholders? What changes do they desire?

What are the effects of quality assurance changes on reliability and validity?

Awarding

Given the changing roles of examiners in an on-screen environment, who should be involved in setting grade cut scores?

How should judgements about different evidence be combined?

If examiners do not mark whole scripts, what is the basis on which they may recommend cut scores? Can judgments about items or groups of items be translated into cut scores?

Should Principal Examiners mark whole scripts, even if other markers do not? If they do not mark whole scripts, can they make judgments about cut scores by looking at re-assembled, marked whole scripts? Does it help if they first mark some answers from every item?

What are Principal Examiners thinking about when considering cut score recommendations? Does the software, computer equipment or Internet interfere with their cognitive processes?

What statistical evidence is useful at an Awarding meeting, and how should it be presented?

What methods of selecting and presenting answers or whole scripts to Awarding meeting participants are best?

What are Awarders thinking about in an Awarding meeting, and does the software, computer equipment or network connection interfere with their cognitive processes?

If Assistant Examiners and Team Leaders are no longer able to make judgements about grade thresholds, will they feel marginalised?

What changes are acceptable to Assistant Examiners, Team Leaders, principals and other personnel involved in Awarding and other stakeholders? What changes do they desire?

What are the effects of Awarding changes on validity and reliability?

Conclusion

Clearly it would be impractical to investigate every possible way in which on-screen marking might be used and all the questions raised above. Previous research findings and business plans must guide the choice of what to consider and set the immediate, mid term and longer term research priorities. As we illustrate above, however, moving to an on-screen marking environment is not a straightforward matter, and many factors need to be investigated in addition to technology issues. In such circumstances a simple transfer of existing practices to an on-screen environment is unlikely to prove viable, and deeper changes must be made that fully take advantage of the benefits that on-screen marking may bring.

Brief summaries of research so far

So far UCLES and its UK trading subsidiary OCR have undertaken two research studies[3]. The first study investigated the centres' experience of two alternative ways of providing machine-readable script identifiers. The other involved an initial exploratory session of on-screen marking with some senior examiners.

The centres' experience

Aim

The aim of the research was to evaluate two approaches to providing machine-readable script identifiers and their impact on centres (schools and colleges).

Background

OCR recruited seven local centres to trial two alternative ways of providing machine-readable script identifiers in the January 2004 examination session. These approaches, described above, were Fully Personalised (FP), where all identification details were pre-printed onto question papers and answer booklets, and Intelligent

[3] We should like to acknowledge the help and support of our colleague David Glover who was OCR’s project manager.

Character Recognition (ICR), where some details were automatically read from candidates' handwriting.

The ICR approach requires centres to make no changes to their current practice. The FP approach, however, requires that in addition to their usual checks centres must ensure that each candidate uses the right pre-personalised answer booklets or question papers. Where pre-personalised answer booklets are used centres must store sufficient of these for every examination taken by every candidate. In the conventional – and ICR – methods no examination or candidate details are pre-printed so any answer booklets may be used. The FP method may therefore require centres to store more stationary.

Three centres used FP and four centres used ICR for all general (i.e. not vocational) OCR examinations taken at their centres in the January 2004 session.

Method

Four methods of data collection were used: researcher observations of two examinations in each centre, incident record sheets completed by each centres' invigilators (proctors), questionnaires completed by invigilators when the January examinations were over, and finally a semi structured interview with each centre's Examination Officer(s) – these are a centre's principal examination administrators.

The first occurring examinations were not observed by the researchers, and invigilators were not asked to complete incident record sheets concerning them. This was so that the evidence collected reflected how the procedures worked after initial "teething" problems were sorted out – OCR staff were available to help centres with teething problems. Observation checklists covered characteristics such as the numbers of invigilators, candidates and examinations taking place in the examination room, how the room was prepared and how long this took, how candidates knew where to sit and how long this took, what instructions invigilators gave candidates, the checks they performed and how any problems were dealt with. Invigilators' incident record sheets asked invigilators to record whether examinations started and ran on time, whether details printed on question papers and answer booklets were correct and any problems connected with getting materials to the right candidates. Invigilators' questionnaires and Examination Officer interviews covered their experiences of working with the method trialled at their centre and their opinions, concerns and suggestions.

Findings

Analysis of data from the incident record sheets and observations showed no consistent differences between the ICR and FP methods.

During interviews, however, Examination Officers made considerably more negative comments about FP than ICR. They also made a few more positive points about FP than about ICR, but this disparity was far smaller than the difference in terms of negative feedback between the two methods.

Invigilators' questionnaire responses also gave more negative feedback about the FP method than for the ICR method. Their positive comments were mixed.

Limitations

The ICR centres tended to be centres with bigger entries than the centres using the FP approach.

More invigilators from centres trialling ICR returned questionnaires than did those from FP centres.

Some of the ICR centres had experience of the FP approach from other Awarding Bodies. Some invigilators and Examination Officers using the ICR approach made comments which were evidently about the FP approach and were presumably based on their personal experiences of other systems or hearsay.

Some patterns in the data did not appear to be a result of the FP or ICR approaches. Observed patterns in the examination halls tended to be determined by the centre and/or the examination, for example, the size of the entry at a centre. The invigilators' questionnaires revealed that FP invigilators were generally less experienced than the ICR invigilators. This is likely to be due to the centres and their procedures.

Some information from different methods of data collection tallied, validating the authenticity of the data.

Conclusion

Given the limitations of the data and evaluation design, firm and unambiguous conclusions cannot be reached. However the FP approach was more negatively received than the ICR approach, particularly in larger centres. With this in mind it is thought that the ICR approach will be more favourably received by centres than the FP approach. An analysis performed by colleagues covered the accuracy of the ICR data capture and checking processes and concluded that these were very good and acceptable, and so UCLES will move forward with this method in accordance with centres' preferences.

The first, exploratory on-screen marking trial

Aims

The main aims of the on-screen marking study reported here were to begin to:

- recognise and include examiners as a stakeholder group within the current on-screen marking programme;
- identify and investigate the underlying causes of any concerns or uncertainty within the examiner population relating to the introduction of on-screen marking;
- develop a clearer picture of the benefits and limitations of on-screen marking from an examiner's perspective;
- collect informed opinion about the features that an on-screen marking system should provide within the UCLES context.

Methods

ETS's Online Scoring Network (OSN) system was used. Four GCE subjects were involved: Mathematics, General Studies, Physics and Chemistry. A few scripts scanned during the January 2004 trial were loaded into OSN. On-screen marking trials were held over four days in Cambridge, with each day devoted to a single subject. For each subject three senior examiners, plus the relevant Chair of Examiners, took part. Data were collected via direct observation, verbal protocols, focus groups and chaired discussions.

Findings

Consulting and involving examiners:

Participants greatly appreciated being consulted and involved at such an early stage and would like to continue to be involved. All examiners, at all levels, are likely to already have heard – sometimes misleading – rumours about on-screen marking and have worries and concerns about it. Examiners suggested that on-screen marking should be developed and introduced in a way that is sensitive to examiners' needs and concerns. They appreciated that there should be an "iterative" approach to development and introduction, with examiners involved throughout. They suggested that arrangements be made to enable Assistant Examiners to try a demo version of the marking system at home as soon as possible. They wanted UCLES to be open and honest about UCLES' plans and aspirations and to consider having a Website where examiners and others can find out about UCLES' on-screen marking programme.

On-screen marking software features:

- Different features will be required depending on the subject, level and item type.
- The examiners sometimes needed to be able to record more than one mark record for a question part, and these must be appropriately labelled. For example, in mathematics, examiners need to record method marks (M), accuracy marks (A) and correct result marks (B) separately, whereas in some science items examiners have to record quality of written communication (QWC) marks separately from "content" marks.
- Current mark schemes sometimes require examiners to award a mark **not** associated with any particular item, but with a group of items. For example, in some science papers an holistic QWC mark is awarded based on a candidate's answers to several questions.
- Examiners are accustomed to annotating candidates scripts with annotations such as E.C.F. (Error Carried Forward) – there are many more. These annotations are determined by the mark scheme, written by the Principal Examiner. Examiners felt that they needed to be able to use these annotations during on-screen marking in order to mark properly.
- Examiners are also accustomed to ticking scripts and wanted a facility to continue to do so.
- Examiners sometimes want to superimpose other subject-specific marks on the image (e.g. an "omission" indicator in Mathematics).
- Some examiners wanted to be to highlight specific portions of the image (and to save these highlights).
- Some examiners wanted to be able to make textual comments linked to a specific (perhaps highlighted) part of the image.

- Some examiners sometimes wanted to make textual comments at a global level (e.g. at the end of a question or group of questions or paper).
- Some examiners wanted the system to automatically tot up sub-marks for a question part (e.g. M and A marks in maths, QWC and content marks in science), or automatically add up ticks when there is a one-to-one correspondence between ticks and marks. If sub-marks are added automatically a confirmation facility might be useful before moving on to the next item.
- In addition to pre-set options for indicating why a script is being referred to a Team Leader examiners wanted to be able to enter textual comments when referring scripts.
- Some examiners are accustomed to putting a problem script or answer to one side until later, then coming back to it, perhaps after consulting a Team Leader for help or looking back to see how they coped with a similar problem previously. They are also used to recording a textual comment about why they have deferred a script for later marking, to jog their memory when they come back to it. They wanted to be able to continue to do this.
- Currently examiners know how much work they have left in a marking season by seeing how many envelopes of unmarked work remain. They would like a similar indicator in an on-screen environment. This information is useful for personal satisfaction and for time planning.
- Some examiners mark item by item as they learn the mark scheme and then mark whole script by whole script when they are familiar with it. Examiners requested that they be able to continue this practice when on-screen marking.
- Examiners wanted to be able to use mouse wheels whilst on-screen marking.

Whole script, section or question level marking?

- Participants expressed a clear preference for marking whole scripts, though on occasion, for example when becoming familiar with a mark scheme, some examiners liked to mark several answers to a question together.
- Participants generally felt that items within a question should **never** be split for separate distribution.
- If some mathematics or science questions are to be separated from the rest of the script for clerical marking then the Principal Examiner (or other paper author if not authored by the Principal Examiner) should decide which ones and where splits should occur.
- Some science papers have Quality of Written Communication marks that extend over several questions. These questions need to be marked together.
- When deciphering a hard to read number or symbol it can be helpful to look at other examples elsewhere in a candidate's script.
- Continually adjusting to different handwriting was a problem for some examiners, particularly if image settings needed to be changed. This problem was most acute for short answers.
- Participants were worried that for papers where candidates answer on separate answer booklets, some work might not get marked if items were apportioned to

different markers. This is because some answers might be misidentified due to candidates labelling them incorrectly or illegibly, and parts of some answers may not all be written together. This latter point may also apply to papers where candidates answer on the question paper if scripts are split into too small chunks, since when candidates wish to change a previous answer they may cross it out and write elsewhere.

- One or two examiners suggested reasons why whole script marking might be less objective than "split" marking.
- Examiners felt their satisfaction would be reduced if they did not mark whole scripts. Some teachers mark to gain insights for their teaching and so would not want to just mark one or two questions.
- Examiners noted that Principal Examiners have to make recommendations about grade thresholds for Awarding and therefore need to mark whole scripts.
- If Team Leaders and Assistant Examiners are no longer in a position to provide consultancy to Principal Examiners about grade threshold recommendations due to not marking whole scripts they may feel their professionalism is diminished.

Examiner recruitment and retention:

Many participants were worried about the prospect of losing older examiners who might feel they lacked the computer skills needed or didn't want to face the stress of change. Participants also speculated that new, younger examiners might be recruited as a result of on-screen marking. UCLES should consider how to attract new blood, and how to accommodate those who cannot or will not adapt to the new system. Could some question papers within a syllabus be marked on screen and some on paper, with examiners assigned to the paper that is marked in the way they prefer?

Practical issues:

Participants felt several practical issues were of great concern to them and other examiners:

- They should not incur personal costs (e.g. Internet Service Provider costs);
- UCLES would be taking a liberty in requiring examiners to provide and use their own equipment for long periods;
- Would sufficient examiners have a suitable computer? If broadband was needed, would it be available and who should pay?
- Home computers were needed by other family members;
- Location of home computers and phone lines greatly restricted where examiners could mark;
- What health and safety and "comfort" issues were there associated with using a computer for long periods?
- Would sufficient examiners have the necessary computer skills or inclination to acquire them?

These concerns were often expressed in the initial expectations session as well as in the post marking focus group and so are likely to be widespread amongst examiners generally. Participants also felt training and technical support would be needed.

Other marking and professional issues:

- All examiners strongly preferred paper mark schemes and question papers. There was little or no demand for on-screen mark schemes or question papers.
- Some examiners were concerned that on-screen marking might reduce the reliability and validity of marking and wanted this checked.
- Examiners greatly valued face to face standardisation meetings and felt these would not work well if everybody was using a computer during them.
- Some participants felt that on-screen marking might be quicker for short answers and longer for long answers. Some other examiners felt that working on-screen for long periods might slow marking down.

Scanning and item identification:

- Variable image quality must generally be avoided, though answers involving maps and graphs may need colour and higher resolution.
- Examiners need to feel sure they have everything a candidate wrote for an answer, otherwise they may worry and may be more lenient through being more inclined to give candidates the benefit of the doubt when answers appear incomplete.

Quality Assurance

Many participants were very concerned that some new ways of monitoring marking quality and examiner performance smacked to them of “Big Brother” and would strike Assistant Examiners similarly. Some participants felt that interspersing “gold standard” scripts (where marks have been pre-decided but kept secret from examiners) was deceitful and would destroy trust and team spirit. Even if examiners knew that a script was a gold standard script before they marked it this would still be unacceptable to some examiners. Randomly and even secretly reviewing marked scripts was far less controversial, as was the idea of sending occasional scripts to many or all examiners, providing they were not pre-marked. One reason put forward for these views was that with gold standard script marking one is looking for deviations from the “correct” marks, but if marks have not been determined in advance one is looking to see whether one agrees with an examiner’s marks.

Examining depends on examiners exercising their professional judgement. Some participants felt that once Team Leaders have conducted their checks and approved Assistant Examiners to begin marking additional monitoring would imply a lack of trust in AEs’ professionalism. They felt that it smacked of trying to catch examiners out and penalising them for occasional slips.

References

Bennett, R. E. (2003) *On-line Assessment and the Comparability of Score Meaning* (ETS RM-03-05), Princeton, NJ: Educational Testing Service.

Newton, P., Whetton, C. Adams, E. Bradshaw, J. and Wong, C. (2001) *An Evaluation of the 2001 New Technologies Pilot*, NFER.

Mead, A. D., and Drasgow, F. (1993) Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, 114(3), 449-458. In Bennett, R. E. (2003) *On-line Assessment and the Comparability of Score Meaning* (ETS RM-03-05), Princeton, NJ: Educational Testing Service.

Powers, D. and Farnum, M. (1997) *Effects of Mode of Presentation on Essay Scores*, (ETS RM -97-08), Princeton, NJ: Educational Testing Service. In Zhang, Y., Powers, D. E., Wright, W. and Morgan, R. (2003) Applying the On-line Scoring Network (OSN) to Advanced Placement Program (AP) Tests. (RR-03-12) Princeton, NJ: Educational Testing Service.

Powers, D. E., Farnum, M., Grant, M., Kubota, M. (1997) *A pilot test of on-line essay scoring* (ETS RM – 97 – 07), Princeton, NJ: Educational Testing Service. In Zhang, Y., Powers, D. E., Wright, W. and Morgan, R. (2003) Applying the On-line Scoring Network (OSN) to Advanced Placement Program (AP) Tests. (RR-03-12) Princeton, NJ: Educational Testing Service.

Powers, D., Kubota, M., Bentley, J. Farnum, M., Swartz, R. and Willard, A. E. (1998) *Qualifying Essay Readers for an On-line Scoring Network* (ETS RM – 98 – 20), Princeton, NJ: Educational Testing Service. In Zhang, Y., Powers, D. E., Wright, W. and Morgan, R. (2003) Applying the On-line Scoring Network (OSN) to Advanced Placement Program (AP) Tests. (RR-03-12) Princeton, NJ: Educational Testing Service.

Sturman, L. and Kispal, A. *To e or not to e? A comparison of electronic marking and paper-based marking.* Paper presented at the 29th International Association for Educational Assessment Conference, 5-10 October 2003, Manchester, UK.

Twing, J. S., Nichols, P. D. and Harrison, I. (2003) *The comparability of Paper-Based and Image-based Marking of a High Stakes, Large Scale Writing Assessment*, Paper presented at the 29th International Association for Educational Assessment Conference, 7 October 2003, Manchester, United Kingdom.

Whetton, C. and Newton, P. (2002) *An evaluation of on-line marking*, Paper presented at the 28th International Association for Educational Assessment Conference, 1-6 September 2002, Hong Kong SAR, China.

Zhang, Y., Powers, D. E., Wright, W. and Morgan, R. (2003) *Applying the On-line Scoring Network (OSN) to Advanced Placement Program (AP) Tests*. (RR-03-12) Princeton, NJ: Educational Testing Service.