Lessons In The Development And Deployment Of Automated IT Skills Accreditation

Stewart Long
School of Information Systems
University of East Anglia
Norwich
NR4 7TJ
UK
Tel: ++44 (0)1603 592607
Email: sl@sys.uea.ac.uk

Abstract

Assessment software is currently being deployed by a UK Examination Board to largely replace human assessment for the accreditation of word processing skills for one of its most popular awards. This paper is intended to outline some of the key lessons learned concerning assessment software *development* and *deployment* which have been instrumental in the project's success.

While commercially produced software tends to employ technology driven solutions which compromise educational quality for development speed, academic research is often theory driven, resulting in incomplete, specialised and unprofessional systems. The CAA system described herein has had to preserve all aspects of assessment quality. This has been achieved through long-term and close collaboration with the Examination Board, and the development of rigorous and innovative solutions. This has included extensive evaluation designed, not only to inform development, but also to build client confidence.

In addition organisational, logistical and culture change issues had to be addressed for CAA to be successfully deployed. This required a delicate balance between integration and innovation.

In conclusion it is argued that technologically attractive methods do not necessarily provide appropriate solutions. These are only achieved for authentic assessment scenarios through long-term and close collaboration with educators.

Keywords

IT skills accreditation, word processing, automated assessment

Introduction

This paper focuses on the use of computer-based assessment methods to replace human examiners in traditional word processing examinations. Earlier work (Dowsing & Long, 1996) carried out at the University of East Anglia under the Teaching and Learning Technology Programme (TLTP) has been applied to automating the assessment of one of the most popular word processing awards delivered by a leading U.K. Examination Board. In the area of practical IT skills assessment, while informal assessment systems are relatively common, CAA methods are rarely used for formal accreditation. This collaborative project, however, has resulted in full integration of practical CAA in a professional examination system. Some of the secrets of the project's success are discussed here with particular references to two key areas. The first concerns the *development* of software that can adequately perform the assessment task currently carried out by human examiners. The second part concerns the *deployment* of the technological solution in the context of a large number of examination centres, multiple yearly intakes and many thousands of examination candidates. The paper concludes with a summary of the successes of the project and the lessons learned.

Background: the target examination system

Examinations may be taken at any one of approximately 800 affiliated exam centres which include Further Education Colleges, Adult Education Centres, and IT training centres. Candidates are required to produce/edit 3 short documents. Instructions are presented on paper showing what alterations to make to the original documents. The sub-tasks involved include inserting, deleting, replacing and moving text, formatting characters, paragraphs and pages, altering the layout of the document, and creating or editing tables. Candidates may use any word processing application to complete the examination tasks in any order and, on completion, printouts of the final documents are sent for marking to designated human examiners.

The human examiners study the final printouts to locate errors which are then classified according to type and context. Assessment criteria dictate how to count errors, for example, once per word or once per examination, and how the total error count relates to the overall classification of distinction, pass or fail. Although the final solutions are well defined, some alternatives are allowed and assessment criteria dictate how to penalise inconsistent use of alternatives. Finally, examiners send marked scripts to the examination body head quarters where standardisation checks take place before results are disseminated.

Lessons in CAA Development

Educational software often makes a trade-off between technical difficulty and educational complexity. Commercially produced software tends to favour technology driven solutions, and risks compromising assessment validity. For example, traditional CAA techniques such as multiple choice questions, or fixed tests of atomic functions, fail to assess authentic skills (Fletcher, 1992; Mager, 1990). At the other extreme, academic research is often theory driven, and programs produced are incomplete and specialised. They lack educational comprehensiveness and fail to deliver software engineering

quality (Gillies, 1991; Self, 1998). What is needed is a problem driven approach (Gillies, 1991) which not only wins client confidence, but also maintains the involvement of the client through close collaboration. This leads to higher quality software and avoids the “not invented here” syndrome which proves the undoing of many off-the-shelf CAA applications (Adman & Warren, 1993). These key lessons are now considered in more detail.

Provide answers to the right questions

Nurture Collaboration

The target activity traditionally performed by human examiners is very sophisticated, and requires input from a large number of knowledge sources. This requires considerable time and effort to understand and ultimately to reproduce. It also requires developers to work closely with scheme experts and human examiners, carrying out Knowledge Acquisition exercises (Luger & Stublefield, 1989) to ensure that CAA performance and knowledge is comprehensive and accurate. Such sophisticated performance cannot realistically be produced using the traditional waterfall model of software development, which describes a linear progression from design to implementation. In such cases an “incremental evolution” (Dutta, 1993) or “life cycle” (Gillies, 1991) model involving a cycle of prototyping, evaluation and modification is preferable. Close collaboration with task experts during this cycle is essential to ensure performance quality and that the development remains driven by the original problem. The involvement of task experts in this way also contributes to the confidence of clients that the system can achieve the appropriate levels of performance.

Find an Appropriate Assessment Model

Where CAA is targeted at an existing examination system, the assessment model chosen must be able to support the same examination tasks and tools. It must allow the same kinds of candidate performance and exhibit the same assessment sensitivity in order to avoid compromising assessment validity. In addition, it must also avoid fixed or hard-wired solutions, in favour of accessibility, so that the system can be easily modified for new examinations or assessment criteria. For example, where possible, knowledge should be separated from control to make it an accessible resource.

Given that current scanning technologies are not sufficiently reliable the solution adopted was to assess the final output as files, rather than paper-based documents. This means that no change was required in the examination setting and sitting, and the same assessment criteria can be applied. The core of the system’s algorithm mirrors the human examiner’s task which is essentially to a) detect errors in candidate solutions and b) classify those errors according to assessment criteria rules so that they can be counted correctly. Detection of potential errors is based on the comparison of candidate solutions to correct or model solutions provided by the examiner using well documented difference algorithms (Miller & Myers, 1985). Artificial Intelligence, rule-based techniques (Luger & Stublefield, 1989; Dutta, 1993) are then used to represent assessment criteria and classify and count errors appropriately. The system is described in detail elsewhere (Dowsing & Long, 1999a; Dowsing & Long, 1999b).

The document difference approach is also attractive from the point of view of system modification as new examinations can be defined simply by providing new model solutions for comparison. This philosophy of adaptability/ openness is maintained throughout the system so that all assessment data, such as error counting criteria and lists of valid alternatives, are contained in files which are external to the central assessment engine. This means that the system is reusable for new examinations, and that changes can be made in the realm of the user, i.e. the examination scheme expert, without the need for special programming expertise. A wide array of output is also available so that all results are fully justified, and this can form the basis of evidence to prove to the client that adequate performance levels have been achieved. It is also be necessary to support an appeals process.

Strive for Quality

Software quality must not be compromised for development time, or cost or technical solutions will ultimately fail to perform adequately. There are two principle dimensions of quality which must be considered.

Educational Quality

In order to be successful Educational Technology must achieve sufficient quality to be accepted by educators. In the case of assessment software there are several important dimensions which must be achieved (Nuttal & Willmot, 1972; Harlen, 1994; Tuijnman & Postlethwaite, 1994).

- Validity is the extent to which a test or examination measures its intended skills or knowledge.
- Reliability is the ability of the assessment system to consistently produce accurate assessment results.
- Accountability means that an assessment result is open to examination and can be justified if queried.

Software Engineering Quality

As well as addressing questions of educational quality any system intended for widespread use must be reliable, open and extensible. Software reliability, as opposed to educational reliability, refers to the robustness of the software and its tendency to crash or cause other systems problems. Openness is closely related to accountability. Educators will not trust black box solutions whose performance is not open to scrutiny. Neither will they use systems which cannot be easily adapted to their own needs, or updated without the aid of a computer programmer.

Some IT skills assessment systems are being developed which allow candidates to demonstrate more authentic skills. Such systems have either been developed by those responsible for delivery of a specific IT course at an academic institution (Kennedy, 1999), or by commercial bodies such as NCC Education Service's Euro PC Test. However, all of them are fixed or opaque. Systems developed by small groups for a specific course or academic institution invariably implement assessment procedures and materials as fixed functions and resources. Some commercially available systems are based on more general architectures but, given the nature of commerce,

these are hidden from the exam administrator/user, and new texts and materials must be purchased from the supplier.

Win client confidence

In order to be used a system must win the confidence of the client/user. A major component of this confidence is based on evidence of adequate system performance over a wide range of data. This evidence is provided by full and continuing evaluation of the assessment system. Evaluation also has an important formative role, informing improvements to the prototype system through iterations of its development cycle. Useful evaluation dimensions identified for the word processing assessment system include:

Overall Indicators of Performance

At the highest level what matters to examination candidates is that they receive the correct final grade or classification, thus the proportion of incorrectly classified scripts is an important measure of the system's correctness. A further way of analysing how an assessment system performs is to calculate the reliability coefficient of the error counts for each candidate. This is a number between 0 and 1 (the higher the better) which indicates how close a score, or error count, produced by a certain assessment system is likely to be to the true score. This measure is commonly cited in the literature concerning standards in education (Nuttal & Willmot, 1972; Harlen, 1994; Tuijnman & Postlethwaite, 1994) and can be calculated in terms of observed, true and error scores and their variances.

Detailed Performance Analysis

Overall performance indicators are useful, but do not provide any detail about performance. In order to do that a comprehensive analysis must be carried out of every error made by the candidates and/or counted by the assessment system.

Evaluation Results

Tests have shown that the word processing assessor incorrectly classifies less than 5% of candidate solutions, which compares favourably with human examiner performance. Similarly, its ability to count errors is comparable to that of human examiners, and it regularly achieves a reliability coefficient of 0.97 and above. It is acknowledged in the assessment standards literature that reliability coefficient scores of over 0.9 are desirable, and that over 0.96 are considered very good. However, overall indicators of performance do not give the whole story. Detailed analyses of human and computer performance have shown that there are certain fundamental differences in their assessment performance. It has been discovered that the CAA system is better at detecting potential errors than human examiners, but that a human examiner, once presented with an assessment issue to resolve, is highly unlikely to misclassify a potential error. These findings make sense given that the computer never gets tired or suffers lapses of attention, but neither can it have access to the human examiners vast resources of domain specific and common sense knowledge.

CAA need not be all or nothing

While early performance evaluation results were promising, it became clear through the cycles of prototyping and evaluation that certain candidate errors were more difficult for the system to assess than others. It was found that, while most candidates' solutions could be assessed accurately by the system, there were a few inaccurate results, usually due to overzealous error detection. In order to provide a mechanism for dealing with these situations, and as a way to allow a more gradual and risk-free route to the deployment of automated assessment, the notion of uncertainty and reassessment was developed. Whenever the degree of uncertainty in the error count is such that the final classification is uncertain the candidate's solution is flagged for reassessment by human examiner. This currently occurs in some 10-15% of cases. As confidence in the system grows the rules can be made more certain and the number of scripts sent for reassessment will diminish. Human examiners are required only to resolve certain assessment issues where the computer is not certain of the result. They do not have to reassess scripts from scratch. This means that all the advantages of computerisation are retained and the new role for human examiners plays to their own particular strengths.

Lessons in CAA Deployment

In spite of the quality of a piece of educational software, it will fail to achieve widespread use if organisational, logistical and culture change issues concerning its deployment are not addressed. From the point of view of the developer, the deployment of educational software may prove to be a more difficult problem to solve than its development. This is especially true where technology is embedded in large or complex systems or organisations as many people, procedures and structures are affected. Whereas during development the developer retains a high degree of control over the project, control of deployment is spread across many parties. The target examination scheme for the automated assessor described here is delivered within a very large organisational structure made up of hundreds of disparate exam centres, a network of human examiners and overall control in the hands of a team of people in a large central office. Again, time and close collaboration have been vital in understanding how best to integrate CAA into such a system. Some key lessons are now presented.

Minimise risk through phased deployment

Phased deployment (Dutta, 1993) has been vital to the success of the project so far for numerous reasons. The examination scheme whose assessment is being automated is very successful in its paper-based form. All risk associated with altering such a scheme must be minimised so that problems can be located and rectified early and with minimum impact. The gradual cultivation of confidence through the stages of development and deployment has ensured continuing support for the project. Just as confidence in software performance must be cultivated, so confidence that it can be successfully integrated is built gradually. Furthermore, the organisation is itself discovering what it wants from automated assessment and how it must change to incorporate it. The ultimate impact and boundaries of any automation process may be unknown at the beginning of the project. Thus phased deployment

provides the opportunity for developers to learn incrementally about interface requirements and organisations to learn what changes are really achievable and desirable through technology, and how employees responsibilities will change.

Once initial testing had demonstrated that the assessment system could achieve adequate assessment performance, the following stages of deployment were planned:

- Parallel pilot with traditional system using ad hoc submission techniques
- Live pilot using new submission technique (limited numbers)
- Live pilot using new submission technique (large numbers)
- Full availability

Balance integration and innovation

Ideally, new technology should be customised to fit into existing procedures and structures, but some re-engineering of these structures may be inevitable in order to accommodate it. In fact, integration of technology may present a welcome opportunity to analyse and improve the traditional system. Thus a key part of the deployment process has been identifying how to integrate CAA into current practices with the least possible disruptions, but always being prepared to seize an opportunity for innovation to improve upon current practices.

At affiliated centres across the U.K. where candidates sit their examinations, it is very important that few additional technological burdens are placed on candidates or administrators. In a competitive economic environment centres and candidates might vote with their feet if CAA introduced new overheads or demands. Ideally, no new software should be required for candidates to carry out their examination tasks. The chosen assessment model means that, as far as candidates are concerned, the same tasks can be performed using the same tools. In the traditional system documents had to be saved as well as printed, so the tools and procedures for preparing data for CAA were already in place. The only new requirements for candidates are that files now have to be saved in a particular format (Rich Text Format) and using more stringent naming conventions. In addition candidates are required to complete a small online form, in place of a traditional paper-based one. This is done under supervision before the exam begins.

Administrators at examination centres have expressed worries that the use of CAA might introduce a new administrative burden associated with preparing lots of files and floppy disks for submission to the examination board, in spite of the fact that overheads associated with preparing paper-based submissions are reduced. An innovative method of file submission has been introduced which helps to minimise this new workload. The examination body has recently introduced a World Wide Web interface with examination centres. The submissions process has now been channelled through this interface.

Although automation promises to reduce administrative overheads, if care is

not taken it can simply introduce new ones. Using direct data exchange between centres and the central examination body has been an important innovation as it means that the central office is not swamped with floppy disks containing submissions. Similarly, a considerable amount of work could be associated with the preparation, validation and management of data for automated assessment. This has been avoided by introducing a highly innovative automated process control and assessment management system. A dedicated workflow system, developed by a third party, has been introduced which automatically picks up electronic submissions to create a database of centres and candidates which drives the various stages of data validation and assessment. Although human monitors can intervene at any time, this essentially means that the entire process from submission to assessment can take place without the need for human intervention. Where candidate solutions are marked for reassessment tasks are automatically generated and sent to examiner in-trays which interface with reassessment tools. Such an innovative approach to data submission and assessment management has required a considerable amount of work, especially as the responsibility of development and management of such systems falls within the remit of several different parties both within and outside the examination board. However, this initial effort will avoid considerable overheads in the long term, and pave the way for similar management of CAA approaches for other examination schemes in the future.

Finally, it has been important to understand how CAA must integrate with other procedures such as moderation, standardisation, awarding systems and quality assurance.

Appreciate Culture Change Requirements and Opportunities

Either as a direct consequence of deploying technology, or as a side effect, certain culture changes may come about, while others may be necessary to ensure success. Management must be aware of the politically sensitive nature of these changes, and be able to focus on their positive aspects. In the case described here culture change is needed at the exam centres to shift the focus to working with files rather than paper-based documents. This is partly a training issue, but it goes beyond that as there is a particular longstanding mindset associated with working predominantly with paper-based output. This must be challenged so that candidates, tutors and administrators fully understand that the evidence on which candidates are judged is no longer a printed document, but the final version of the file they submit. Rather than being a problem, this should be viewed as an opportunity to promote best practice and encourage useful workplace skills.

At present CAA is being deployed for only one of many examination schemes delivered by the examination body, with others under development. Even so, some members of staff will experience a change in the quantity and nature of their work as a consequence of CAA. This may simply mean a move from paper-based administration to electronic administration, or it may provide an opportunity to have a closer involvement in educational matters, such as examination design or standardisation. In any case, all affected groups should be involved and kept informed of progress as early as possible. Many people,

especially where well-established traditional systems are concerned, naturally view new approaches with some suspicion. They must be convinced that possible changes to job descriptions will actually make their work more interesting. This is especially true as educational/assessment experts, not IT experts, are best suited to maintaining educational system.

Some of the key ways found to encourage take-up both at examination centres, and among examination board staff, include

- incentives for centres which participate in trials
- regular visits and training days for centres taking part in trials
- regular updates for staff on project developments
- involvement of staff in design decisions
- regular staff training sessions

The deployment of CAA also promises to change aspects of examination body culture in other ways. For example, automated assessment is much quicker than the conventional paper-based system, and also provides a great deal more readily accessible information concerning candidate performance. This means that the potential exists for analysis of the additional data and the incorporation of the results into re-runs of the assessment system, all within the time-scale of a single conventional assessment round. Processes such as exam paper modification, standardisation and plagiarism checks can be expanded in this way, and applied to large numbers of solutions in a short time. The potential therefore exists to radically improve some of the traditional Quality Assurance procedures employed by the examination body.

Conclusions

The work described here has lead to the development of automated assessment technology with excellent performance for a well known professional word processing examination. It has also brought about the fruitful reappraisal of current assessment criteria and procedures. The system has been piloted and phased deployment is nearing completion, culminating in full availability from September 1999. Further collaboration on additional examination schemes is also well in advance. These successes have hinged upon the application of the lessons learned throughout the course of the project.

The key lessons concerning CAA development are that computer-based solutions should provide answers to the right questions, they must strive for quality, and ultimately that they will not be taken up unless they win client confidence. It has also been argued that CAA should be deployed using a phased strategy which minimises deployment risk, that integration and innovation must be balanced and that culture change requirements and opportunities need to be appreciated.

CAA cannot be considered a quick fix. Only through long term collaboration will systems be produced with sufficient coverage, validity and reliability to inspire the confidence of educators, and without such confidence systems will not be taken up. Even then, continued efforts in planning and collaboration

are required to ensure that deployment is a success.

Acknowledgements

The author wishes to acknowledge the help and support of Oxford, Cambridge and RSA Examinations (OCR) in all aspects of the work described in this paper.

References

Adman, P. and Warren, L. (1993) *Paper 8: Frames of mind.* in Bull, J. (editor) Workshop on Assessment of Learning in Higher Education, pages 41-46. TLTP project ALTER.

Dowsing, R.D., Long, S. and Sleep, M.R. (1996) *The CATS word processing skills assessor.* Active Learning 4, July 1996. CTISS Publications, University of Oxford.

Dowsing, R.D. and Long, S. (May, 1999a) *The Algorithmic Basis for IT Skills Automated Assessment.* Proceedings of Computers in Advanced Technology conference, CATE'99, (IASTED sponsors) Cherry Hill, New Jersey, USA.

Dowsing, R.D. and Long, S. (July, 1999b) *An Evaluation of the Impact of AI Techniques on the Automated Assessment of Word Processing Skills*. Proceedings of International Conference on Artificial Intelligence in Education, AIED'99, (IJAIED sponsors) Le Mans, France. IOS Press.

Dutta, S. (1993). *Knowledge Processing and Applied Artificial Intelligence*. Butterworth-Heinemann.

Fletcher, S. (1992). *Competence-based Assessment Techniques*. Kogan Page.

Gillies, A.C. (1991). *The Integration of Expert Systems into Mainstream Software*. Chapman & Hall Computing, London.

Harlen, W. (editor)(1994). *Enhancing Quality in Assessment*. Paul Chapman Publishing Ltd, London.

Kennedy , G.J. (1999). *Automated Scoring of Practical Tests in an Introductory Course in Information Technology.* Proceedings of Computers in Advanced Technology conference, CATE'99, (IASTED sponsors) Cherry Hill, New Jersey, USA.

Lugar, G.F. and Stubblefield, W.A. editors (1989). *Artificial Intelligence: The design of Expert Systems*. Benjamin/Cummings.

Mager, R. F. (1990). Making Instruction Work. Kogan Page.

Miller, W. and Myers, E.W. (1985). *A file comparison program*. Software - Practice and Experience, 15:1025-1040.

Nuttall, D.L. and Wilmott, A.S. (1972). *British Examinations: Techniques and*

Analysis. National foundation for Educational Research.

Self, J. (1998). *Grounded in reality: The infiltration of AI into practical educational systems*. From IEE Colloquium on Artificial Intelligence in Educational Software, organised by Professional Group A4 (Artificial Intelligence), held at Savoy Place, London on Friday 12 June 1998. Digest No: 98/313.

Tuijnman, A.C. and Postlethwaite, N.T. (editors)(1994). *Monitoring the Standards of Education*. Pergamon, Oxford.

Warm, J.S. (editor)(1984). *Sustained Attention in Human Performance*. John Wiley & Sons, Chichester.