Knowledge-Based Systems for Marking Professional IT Skills Examinations

Stewart Long
Roy Dowsing
The University of East Anglia
sl@sys.uea.ac.uk

Patrick Craven
OCR Examinations

Abstract

As the use of Information Technology (IT) increases, so does the need for accreditation of IT skills. Of the many Computer-Based Assessment (CBA) systems which claim to assess such skills, most are based on approaches such as multiple choice questions or atomic functions tests within controlled environments. In contrast, most professional qualifications in the U.K., assessed by human examiners, focus on the output of authentic skills, that is, complete documents produced during realistic interactions with industry standard software. In order to automate the assessment of such examinations the expertise and knowledge of human examiner must be represented and the authentic nature of the assessment tasks retained. This paper describes a suite of Knowledge-based CBA systems for IT Skills, developed at the University of East Anglia, which have been deployed by a leading U.K. examination body to replace human markers for several of its flagship IT awards.

1 Introduction

This paper describes the application of Knowledge-Based System techniques to a set of real world problems in an educational setting. The domain area is the assessment of Information Technology (IT) skills, and the applications described have been successfully deployed by a leading U.K. awarding body as an alternative to human examiners for a number of their well-established, market leading awards in word processing, spreadsheets, databases and other IT skills areas.

Most well established professional qualifications in IT employ “performance-based” or “authentic” assessment [1, 2] which “mirrors real-life settings through relevant student tasks that emphasise contextualised and complex problems”[3]. In addition, assessment is carried out by human examiners who use their expertise and experience to apply sophisticated assessment criteria to the candidates’ work in order to grade their overall performance. While potential benefits, including reduced costs, increased speed and availability, and improved assessment accuracy, have been associated with automated assessment, standard Computer-Based
Assessment (CBA) models are unable to support authentic skills assessment, or sophisticated assessment criteria. Some examination schemes have emerged in recent years which have been specially designed to facilitate automated assessment [4, 5], but even these schemes are predominantly assessed by human examiners.

Early work in Artificial Intelligence in Education (AIED), typified by Intelligent Tutoring Systems (ITS) such as Clancey’s GUIDON [6], and Anderson’s LISP Tutor [7], promised to offer a deeper understanding of student skill and knowledge, but the number of AIED systems currently in use outside the laboratory remains very small. Several reasons for this poor take-up have been proposed, including the tendency for theory driven development [8], and the lack of long term industrial investment needed to develop useful systems [9]. Real world educational applications where Artificial Intelligence (AI) techniques are being used tend to be developed bottom up, to solve particular educational problems, with AI techniques introduced only as system behaviour requires them [8].

The systems described in this paper, unlike traditional CBA approaches, have been designed to preserve the authentic assessment model. Knowledge-Based Systems techniques have been introduced, where appropriate, to reproduce the salient aspects of human examiner performance. In this way the computerisation of an authentic real world assessment scenario has successfully been achieved.

This paper firstly describes the problem area, including the professional vocational assessment model and the key aspects of human examiner performance and knowledge. It then outlines the techniques used to automate the target assessment activity and describes an example system, for the assessment of word processing skills. Some of the aspects of application building and deployment which have been key to successful uptake of the assessment system are listed and the paper concludes with an outline of the potential and realised benefits of automation.

2 Problem Description

The problem addressed by the project under discussion was how to automate the assessment process for traditional examinations schemes which feature authentic IT skills assessment performed by expert human examiners. Work initially focussed on word processing skills, and was then applied to the use of spreadsheets, databases, and more recently, web page creation, and charts and graphs. Work took place in collaboration with a leading examinations body with the aim of automating the assessment of some of their leading IT awards. Two areas were identified as crucial to the success of the project: The preservation of the performance-based examination model and the modeling of the key aspects of expert examiner performance and knowledge in order to support the same sophisticated assessment criteria. These areas are now considered in more detail.

2.1 Preservation of the Assessment Model

The assessment systems described in this paper target a number of IT qualifications offered by Oxford, Cambridge and RSA examinations (OCR), one of the UK’s leading examination boards. While the target skills and applications vary across the
schemes, all the schemes employ an authentic assessment model with a number of features in common which any CBA system must support.

Examinations are taken at affiliated centres across the U.K. Candidates to carry out a range of tasks to create and edit a series of simple, but realistic, business documents, spreadsheets, etc. There is considerable freedom in terms of deciding task ordering, which specific methods to use, and to detect and correct errors during the course of the test. Assessment focuses on direct evidence of practical skill, rather than knowledge, and, in particular, on the correctness (of both content and presentation) of the final product, as opposed to the assessment of methods employed. Hence, in the traditional system final solution documents are printed and sent to human examiners. Candidates are allowed to use any available software application capable of providing the necessary functionality and suitable output (via a printer). This is important for two reasons. Firstly, the target examination schemes are taken at many hundreds of examination centres across the country, with a great variety of available tools and resources. Being non-prescriptive about software means that no candidate or examination centre is excluded as examinations can be taken using existing machines and software. Secondly, from the educational and vocational point of view, it is important that the award be seen as the certification of generic IT skills which are not software specific. The introduction of computer-based assessment should not therefore place additional software or hardware restrictions on candidates. At periodic intervals new examination papers are made available within the same scheme. On occasion it may also be the case that changes are required to assessment criteria and rules. In addition, assessment is transparent and supports a number of quality control checks including the candidates right to appeal against an assessment result. Recently formal standards and requirements for vocational assessment in the UK have been implemented by the Quality Control Agency (QCA) as part of the National Qualifications Framework (NQF)[10].

2.2 Expert Human Examiner Performance

At the core of the assessment process to be automated is the human examiner. The role of the examiner is to apply the published assessment criteria to candidate solution documents to detect, interpret and count errors, and determine a final grade (distinction/pass/fail) for each candidate based on the final error total. Thus the principle components of human examiner performance are the detection of errors and their classification (interpretation according to criteria so that they can be counted appropriately).

2.2.1 Error Detection

The human examiner considers each task described in the examination paper and looks for evidence of its performance in the candidate solution. Once the examiner has established which part of the candidate's solution corresponds to the current task it is necessary to establish whether it deviates in any relevant way from the correct, or model, solution. Although "worked examples" are produced as part of standard examination preparation, examiners tend not to use an explicit model answer for reference. Instead they use only the exam paper, which describes the
tasks to be carried out, and the initial document, with which the examination begins, to construct in their own minds the relevant component of the model solution, and any valid alternative components, with which to compare the candidate solution. Error detection is principally perceptual, requiring sustained attention to detect infrequent deviations in relatively similar artifacts. However, examiners also bring a certain amount of knowledge to bear. For example, they employ procedural knowledge of the target skill area in order to decide the most intuitive match, and knowledge of what constitutes a valid alternative, in order to allow alternatively spelled words or acceptable but different formats. This may come from published sources, standardisation meetings, or from their own grasp of semantics, grammar, and the target skill area.

2.2.2 Error Classification

Once a potential error has been detected the examiner must classify it according to published assessment criteria. This is the most obviously knowledge-based activity which the examiners undertake. In order to do this the potential error must be interpreted by the examiner using their procedural knowledge and experience of observing candidate errors. This is a context-free classification but examiners must then determine the relevant context which may also affect the way in which the error is counted. The context might concern the task which the examiner infers that the candidate was carrying out when the error occurred, or the kind of document component in which the error appears. At the core of the examiners' knowledge are the mappings for context free error classification, the ability to recognise the relevant contextual information and the assessment criteria rules which link error type and context to overall error categories and error counting methods. While much of this knowledge is explicit and can largely be found in published materials, a sizable amount of tacit and common sense knowledge is required in order to fill in gaps in the assessment criteria which are necessarily under defined, given the vast number of possible candidate solutions.

2.2.3 Suitability of Standard CBA for the Target Assessment Model

Most of the CBA Systems seen today are based on Multiple Choice Questions (MCQs), or some derivative of the question format [11, 12]. Clearly, they cannot support the authentic assessment model necessary for the target schemes. There are also many systems which claim to assess practical IT skills. However, these are almost always based on the "function test" model. This involves a series of atomic, fine grain tests to see if a candidate can activate individual functions, such as clicking the "embolden" button. Assessment merely indicates if each task has been achieved or not. This view of skill as a series of low level, atomic functions does not agree with the authentic skills model which employs sophisticated assessment criteria to determine if all the evidence identifies appropriate levels of skill or competence. Furthermore, function test systems need to control the candidates' interaction with the application during the examination, hence the use of simulations, macro programs, or other methods for plugging into the event loop while the candidate is actually taking the examination. This requires that candidates use specific software installed on their machines especially for the assessment. This also means that such systems are application specific. Thus,
functions tests are also unable to support the target authentic assessment model. The challenge of automating IT skills assessment is to reproduce the assessment model and then to simulate the sensitivity of the human examiner to different error types and contexts and to adapt error counting procedures accordingly.

3 Application Description

Approaches for preserving the performance-based examination model, and the modeling of the core aspects of expert examiner performance, that is, error detection and error classification, are now described. This is followed by a description of the key components of the word processing assessment system.

3.1 The Automated Assessment Model

The notion of the assessment of completed output files as an alternative to standard CBA approaches for word processing, and other IT skills, assessment was identified in an earlier project [13]. It was realised that examinations whose subparts lead to reasonably well defined goal documents could be assessed automatically by comparing candidate solution files to predefined goal documents to detect differences which, in the earliest systems, equated to errors. In particular, for word processing tests, applying comparison procedures to Rich Text Format (RTF) files [14] was identified as a means of making assessment application independent, given that RTF can be produced by most modern word processors. For other IT skills common output formats are less obvious, for example, most spreadsheets can save in Excel format and web page creation software can save in HTML. Occasionally file conversion software is required to transform from one format to another. This approach supports the fundamental aspects of the professional authentic assessment model described above and allows candidates to take examinations in exactly the same way as the target schemes, with the exception that saved files, rather than printed documents, are submitted.

3.2 Error Detection in Results Files

Difference algorithms [15, 16] have been identified as a technical basis for error detection. They calculate the minimum edit distance between two sequences of tokens, that is, the smallest number of edit operations for the given token size needed to convert one sequence into the other, and, as a by-product, a list of low level edit instructions (in terms of inserts and deletes of tokens) which will reduce the distance to zero. Thus, difference algorithms can be used to find areas of two documents which do not match, thereby detecting potential errors. They can also determine which document components match so that, for example, their formats may be compared. Resultant difference listings can form the basis of error reports for formative and summative assessment, and for justifying system conclusions.

The flat difference comparison of word tokens employed in the early word processing systems was found to provide difference information which was too homogeneous and difficult to analyse. Difference analysis is now structured in order to filter differences according to type and ease additional error interpretation. Difference routines have been modified in a number of other ways in order to
overcome their simplistic heuristic approach and make their results more intuitive, in keeping with human examiners’ error detection methods. These modifications will be discussed in a future paper. For other IT areas the same principal is employed for error detection. For example, spreadsheet files, are comprised of sets of words, numbers and formulae which are grouped into cells, and cells are organised in rows and columns. Difference algorithms have been modified to determine the correspondence between cells in the model and candidate answer.

3.3 Knowledge-Based Error Classification

Differences between files may or may not indicate actual errors, given that they may correspond to valid alternatives, or differences which can be ignored. Where candidate-goal differences do signify error, sophisticated assessment criteria must be employed to ascertain what sort of error has been incurred, and hence how to count it. This stage of the assessment process has been identified as the error classification stage in human examiner performance, and is the most obviously knowledge-based activity. At the highest level a rule-base approach has been used to model published assessment criteria, including tacit examiner knowledge gleaned through Knowledge Acquisition [17]. Thus rules for representing assessment criteria for the target examination have the following structure:

3.3.1 Pattern component:
- Difference operator - The context-free description of the difference between the candidate and goal documents.
- Context – The context, in terms of document or examination task structure, within which the difference occurs.

3.3.2 Action component:
- Counting method - This describes how to count the error if the pattern component of the rule is matched, for example, once per word or per document.
- Error Classification - This is the high level interpretation of the error and matches high level error classifications contained in the published assessment criteria of the target examination.
- Priority - As wildcards may appear in pattern fields of rules it is possible that more than one rule may be applicable at any one time. The priority enables the specification of rules in order of preference for “conflict resolution” [17].
- Certainty Factor – for schemes where reassessment is used a simple certainty factor is attached to rules (1 = certain, 0 = probable) in order to allow the calculation of maximum and minimum final fault counts. Where these cross final grade divides the candidates is flagged for reassessment.

3.4 Description of the Word Processing Assessor

The architecture of the automated word processing assessor is shown in Figure 1 and each stage of assessment is described below. Similar systems have been built for other I.T. skill areas.
3.4.1 Assessment Stage 0: Pre-Processing

Files are initially translated from RTF and a series of low level pre-processing operations are carried out. For example, comparisons are carried out to detect white space differences which may effect the way the document is tokenised into words.

3.4.2 Assessment Stage 1: Structured Comparison

Potential errors are detected by the application of modified difference algorithms. Comparison is structured so that differences are grouped by type, such as word accuracy, case usage, white space and punctuation usage, text misplacements and various different formatting levels. Information on the consistency of use of sets of alternative words and formats is also gathered.

3.4.3 Assessment Stage 2: Context-Free Difference Interpretation

Most of the assessment evidence contained in the difference scripts provided by the initial error detection stage of assessment is described at a low level. For example, the textual differences are described using only three operators; insert, delete and replace. A post-comparison difference interpretation stage has been implemented.
in order to further interpret differences. Rules for mapping low level difference patterns to higher level context-free descriptions are declared externally.

3.4.4 Assessment Stage 3: Context-Sensitive Error Classification

The high level assessment criteria rules for counting and classifying errors match, not only context-free descriptions of the errors, but also the context in which they occur. Context is established through the application of regions to the goal documents. This is carried out as part of the set up process for each new examination paper. Files are maintained, as a resource separate from the assessment engine, which describe the type and location of each region identified in each goal document. These regions are applied to the corresponding components of the candidate documents through the comparison (error detection) stage of assessment.

The following is a sample rule for a word processing examination:

If difference type is MISSING PARAGRAPH and region is TABLE, error type is PARAGRAPH ERROR (3.2), count type is COUNT MAXIMUM OF ONCE PER EXAMINATION, priority is 1

The KA exercise and the testing cycle has focused considerable effort on developing a comprehensive resource of such rules for the target examination schemes and rule-bases typically contain around 150 of these rules.

3.4.5 Assessment Stage 4: Final Result Determination

The final result of summative assessment schemes is a classification which provides an overall evaluation of a candidate’s skill or knowledge level. If CBA is required to determine this final classification at the time of assessment, the classification itself is the principal output of the assessor. Results categories and error tolerances can be represented in an external resource for additional flexibility. However, in the case of the usage scenarios described here, candidate results are stored in a database until awarding processes are set in motion. Thus, the principal output of assessment is the value, or values, required to determine the final classification.

3.5 Assessor Output - Making Results Transparent

As well as calculating an overall result, CBA must support stringent Quality Assurance measures and an appeals process. Expert Systems justify their results by allowing the user access to their reasoning process, that is, the series of rules activated in the search for a solution to the task at hand [18]. Similarly, in order to justify results the CBA systems described here show not only the range, error count and error type for each error detected, but also identify which rule was used to count and classify the error. Two principal methods have been implemented for outputting detailed explanations of assessment results. The first is a text-based report which shows which sections of the document are in error and gives the values of each field in the error counting rule used to classify and count the errors.
The second method attaches the same information to comments in an appropriate application interface via macro programs, so results can be viewed as annotations to the candidates' documents. The word processing assessor can use Microsoft Word® for this purpose, while spreadsheet and database output can be displayed using Microsoft Excel® files.

3.6 Uncertainty and Reassessment

Some schemes targeted have a long turnaround period as solutions must be sent to human examiners from examination centres and returned to the central office before results can be issued. Given this, and the speed of automated assessment, it has been possible to introduce an additional layer of safety and quality control in the use of the CBA. System evaluation showed 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 assessment issues where the computer is not certain of the result. They do not have to reassess submissions 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 [19].

The notion of uncertainty is incorporated into the system by means of a simple certainty measure [20], attached to every error counting rule, indicating whether the rule is certain or uncertain. This essentially allows a maximum and minimum error count to be maintained. If these counts fall into different final categories (distinction, pass, fail) then the final result is uncertain and the candidate solution is marked for reassessment by a human examiner. The size of the rule base is also increased as some error counting rules can have different certainty values in different contexts. This use of uncertainty allows the system to qualify its final judgement about the examination candidate.

4 Application Building/Deployment – Keys to Success

4.1 Research/Industry Collaboration

The relative scarcity of real world systems employing Artificial Intelligence in Education has partially been blamed on a lack of long term funding [9]. A long-term commitment from this project's industrial backers has been vital to its success. This has supported approximately 8 man-years of effort over a five year period from an initial research question – can computers perform authentic IT skills assessment to a standard comparable to that of human examiners – to full
availability of automated assessment for a number of schemes. In addition to financial backing close collaboration with scheme experts and access to resources has been fundamental throughout initial Knowledge Acquisition and subsequent system development and evaluation. Furthermore, such sophisticated performance cannot realistically be produced using the traditional waterfall model of software development, which describes a linear progression from design to implementation. An "incremental evolution" [20] or "life cycle" [21] model involving a cycle of prototyping, evaluation and modification has been employed instead. The involvement of task experts has also contributed to the confidence of clients that the system can achieve the appropriate levels of performance.

4.2 Installation and Deployment Issues

The work carried out at UEA has focused primarily on the development of systems capable of supporting the core assessment models and performance. However, organisational, logistical and culture change issues concerning their deployment must be addressed in order for software to be taken up.

The target examination schemes are delivered at hundreds of disparate exam centres across the U.K. where it is very important that few additional technological burdens be placed on candidates or administrators. The only new requirements for candidates introduced by CBA are that files now have to be saved in a particular format (for example, RTF for word processing schemes) and using more stringent naming conventions. In fact some schemes allow submissions in any format with file translation being performed centrally. 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. In addition, administrators at examination centres have expressed worries that the use of CBA might introduce a new administrative burden associated with preparing 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, through which the submissions process has now been channeled.

Automated assessment, itself, is managed from a central office. A considerable amount of work could be associated with the preparation, validation and management of assessment data. This has been avoided by introducing an innovative automated 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, all of which is controlled at the examination board’s central office. 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 has led to a working automated assessment management model which has now been successfully adapted to a number of different examination schemes.

4.3 Knowledge Maintenance and Adaptation

New examination papers are created periodically, with a frequency ranging from one new paper per year to one new paper at seven specific points throughout the year, depending on the scheme. These new papers are designed to comply with setting specifications so that, although data changes, the same skills are assessed. The distinction between permanent and examination paper specific knowledge for automated assessment has been identified so that only information specific to the new examination papers needs to be declared for each paper, while permanent knowledge, such as rules representing assessment criteria, need not be changed. Thus, in order to prepare a new examination paper, new goal documents, context regions and alternative word lists must be declared. A number of tools have been developed in order to aid users with this task, including tools based in Microsoft Word and Excel which allow text to be selected and regions applied without need for manual word counting or set-up file creation.

4.4 Training and Knowledge Transfer

In order for CBA to be taken up in the target examination systems a number of knowledge transfer and training issues have been addressed. The examination board has organised presentations and training days for centres to make them familiar with new submission procedures. At the examination board staff members have been retrained to perform tasks such as examination setup for CBA, quality assurance checks and maintaining and monitoring workflow and submissions. A number of tools have been developed to ease preparation of materials, and the preparation, running and maintenance of the automated assessment process is now largely carried out entirely by examination body staff. However, there remains a consultancy role for the developers who, given their role as Knowledge Engineers, have developed a deep understanding of the relationship between the mechanics of the CBA system, the external resources for CBA, and the assessment schemes themselves. This kind of in-depth, multi-disciplinary knowledge can only be arrived at over time.

5 Application Benefits

A number of potential benefits of automating the assessment of its existing I.T. awards were identified by the examination board sponsoring the work at the outset of the collaboration.
5.1 Examination board profile
In an increasingly competitive assessment market, it is important for the organisation to position itself at the cutting edge of assessment research. Previously, this was attempted by sponsoring and piloting an I.T. skills assessor based on functions tests, but that system failed to be taken up due to the limitations of its assessment model. Collaboration with UEA on the automated assessment project has successfully positioned the examination board at the forefront of computer-based assessment.

5.2 Cost Reduction
For schemes where large numbers of human examiners are paid to mark examination solutions, the potential exists to reduce the examination board’s costs by automating the assessment process. In schemes where assessment is carried out locally potential cost reductions will be seen at examination centres. However, the advantages may also be felt at the examination board as test centres may promote their schemes above competing schemes, if assessment overheads are lower.

5.3 Assessment speed
Increased speed of assessment is another potential benefit of CBA. The systems described here are, in fact, capable of assessing many candidates in a short space of time – for example, considerably faster than one candidate per second for word processing tests, whereas, it can take an experienced human examiner between 5 and 10 minutes to assess a single candidate. This means that for examination schemes which are taken at discrete times and sent to human examiners for assessment, an entire in-take can be assessed several times by computer and reductions still achieved in the turn-around period between submissions and results dissemination. Where the hybrid assessment system is employed which has additional reassessment by human examiners, a small enough sample of the in-take must be reassessed in order to retain these speed advantages. This is currently the case as approximately 10% of submissions are generally being flagged as probable, allowing reassessment to be performed on a single day, in-house. For locally assessed schemes, a guarantee has been made by the examination board that results will be returned within 48 hours of a batch submission from a centre. In reality the turn around period is much shorter. Anecdotally, email notification of results have occasionally been received at centres as soon as three minutes after submission.
Some additional benefits of automated assessment have been realised even though their potential was not necessarily envisaged at the outset of the project.

5.4 Improved assessment accuracy
Initially, human examiner accuracy was seen as the standard which must be achieved in order to make automation of the target scheme viable. In fact, it has been found that, with good examination setup, and after thorough setup testing, the CBA system can match or improve on human examiner accuracy levels, especially being less likely to fail to detect potential errors [19, 22]. The use of the hybrid assessment system can make assessment results more accurate still.
5.5 Applications of electronic assessment data

CBA provides a large amount of data which is in machine readable form, and thus easy to store and manipulate. At the highest level the CBA returns overall results (in the form of fault counts) which can be stored in a database with candidate details. More detailed information about the frequency, location and type of errors incurred by every candidate is also available. There are a number of possible applications of such data, which have the potential of providing a high degree of benefit. For example, standardisation takes place after a small number of candidates have been assessed by human examiners. Examiners report any issues they have identified, for example, concerning ambiguous examination instructions causing over-frequent candidate errors, or previously unconsidered alternative spellings and formats. A standardisation meeting takes place which results in a message to all examiners offering advice or special guidelines in the light of the issues identified. This process is time consuming and is unlikely to capture all issues, given that time only allows for the assessment of a small proportion of candidate scripts before standardisation. Given the speed of the automated assessment process, and the large amount of data available, an entire submission can realistically be assessed via the CBA/hybrid route and the resultant statistics used to inform standardisation, while still allowing time for assessment to be repeated if changes to assessment resources are required. Tools are under development to do this. The data made available through automated assessment could also be used to carry out plagiarism checks, for example, flagging candidates from the same test centres who incur the same or very similar error patterns. Checks could be made across examination centres to identify possible teaching failings, and similar data could inform the examination paper design process.

6 Concluding Remarks

This paper has presented a case study of a successful application of AI techniques to a real world problem, the computerisation of authentic assessment of IT skills. Unlike most AIED systems, the applications produced have left the laboratory and are now in full use in the real world. Many CBA systems fail to be taken up because, rather than attempting to answer the difficult questions posed by educationalists, they provide answers to their own, simpler questions. This project has set out, from the outset, to develop automated assessment tools for real examinations based on authentic assessment models, not to assume that examinations bodies will drop their proven methods for some simpler, easier to automate assessment model. This has been achieved in a number of ways. Firstly, the model of assessment employed, that is, the assessment of completed output files, without the need to prescribe or control the software employed, offers the possibility of supporting authentic assessment. Secondly, the document difference approach can provide evidence for assessment of the kind available to human examiners. Thirdly, a concerted knowledge engineering effort has taken place in order to reproduce human examiner performance and the sophisticated target assessment criteria. Thorough testing and close collaboration with scheme experts has been important in this regard. Lastly, deployment issues have been taken seriously and dedicated and innovative solutions found for file submissions and
assessment process management. To the knowledge of the authors, the CBA systems described in this paper are the only systems currently being used for, or even capable of, assessing pre-existing, that is, not especially designed, vocational IT awards based on an authentic assessment model.

References
