THE DEVELOPMENT AND APPLICATION OF AN EXPERT MARKER SYSTEM TO SUPPORT 'HIGH STAKES' EXAMINATIONS

Patrick Craven, Peter West, and Stewart Long

The Development and Application of an Expert Marker System to Support ‘High Stakes’ Examinations

Patrick Craven
Peter West OCR Examinations
Westwood Way
Coventry
CV4 8HS

Stewart Long
University of East Anglia

Craven.p@ocr.org.uk

Abstract

The OCR CLAIT scheme was first established in the early 1980’s and has been used as the benchmark assessment for vocational IT skills ever since. It currently attracts over 300,000 entries per year across 9 different application areas. This equates to over 1m scripts per year, all of which require marking, moderation and processing to award of result. Over the last 5 years OCR has worked with the University of East Anglia to develop an automated **‘expert’** marking system for **four** of the application areas within CLAIT (Core, Word Processing, Spreadsheet, Database). This system is now operational and during the first pilot phase successfully supported 30 Centres and processed over 2000 script submissions.

From initial concept the system set out to extend the scope of computer assisted assessment beyond simple skill tests as it was not desirable to reduce the CLAIT assessment to series of atomised functions tests or to place software simulations within testing Centres (over 3500 in the UK). It was also important that the candidate experience of the assessment remained relatively untouched and so an off-line delivery/on-line processing solution was proposed.

This paper provides an overview of the modular approach taken by OCR and reflects on the successes and challenges faced during implementation. The presentation will provide a demonstration of the system with a brief explanation of the intelligent workflow processing and the application of expert marking rules to ensure valid and reliable assessment outcomes.

Introduction

OCR Examinations is a national Awarding Body in the United Kingdom with responsibility for the delivery of ‘high stakes’ assessment on a national scale. Since the early 1980s the IT suite of qualifications administered by the board have been recognised as a benchmark of vocational competence. The qualifications attract around 900k candidates per annum and involve the use of hundreds of expert examiners/markers who are given extensive training to perform their duties to the expected standard. It is accepted that the ‘credibility’ of any given qualification is largely attributed to the collective confidence that the stakeholders will have in the assessment processes that are seen to support the award. Such confidence is also critical to the future ‘trade-in currency’ of the qualification and so is of paramount importance to the Awarding Body. In the face of increasing pressure on the examination system OCR Examinations embarked on a research project with the University of East Anglia to establish whether the qualities of an expert examiner/marker could be captured in a computerised system. This paper gives an account of how such a system has been implemented and reflects on the challenges and successes that the team encountered.

On August 13th 1999 an article appeared in the Times Educational Supplement concerning the reliability and “fairness” of the marking of A level (university entrance) examinations in English Literature in the UK. An experienced examiner was questioning the reliability of examination results, given the large marking load and tight time restrictions faced by examiners, in addition to their normal work. A response was included from a leading figure from one of the most respected examination boards. Naturally, in his position, he stressed the reliability of results based on a system of random checking and re-marking, comparisons to forecast results, and a regulatory body which compares standards across examinations boards. Such discussions are a perennial event. The official line is that the marking of examination board qualifications is as reliable as it is possible to be, but occasionally a dissenting voice is heard, often from within the ranks of examiners themselves. This opposing view maintains that examiners, being only human, and given the demands placed on them, are inevitably prone to error which must be reflected in the accuracy of the results they produce. These concerns resurfaced yet again during the summer of 2001 with examiner overload cited as a cause for many of the difficulties.

The assessment of IT examinations, where solutions are relatively well defined, seems clearly less susceptible to inconsistency and inter-examiner differences in interpretation than do English literature examinations, but IT examiners/markers are also required to assess large numbers of solutions in a relatively short time for relatively little reward. Furthermore, they are required to apply complex, and under defined assessment criteria to detect, count and classify errors. This is by no means a trivial task and one which seems likely to be error prone. This is borne out by empirical studies of human performance in similar tasks. However, empirical results concerning examiner reliability for IT assessment itself are not widely available. This paper represents an attempt to rectify this situation and identify associated issues related to the automation of a ‘human’ process. It presents a model of human examiner activity, derived through the automation of the assessment process, and predicts examiner error patterns, based on findings from the literature concerning the cognitive processes underpinning the key examiner activities. These observations are demonstrated in a study of the performance of human markers and the automated system marking a number of authentic solutions. The paper also reflects on the challenges associated with integrating innovative solutions within traditional processes.

In addition to the study of expert examiner performance the study also considers the qualities of a traditional examination process and what issues should be addressed to automate such a service. It is clear from the study that human processes are very forgiving and allow for minor administrative errors, as the collective 'common sense' of administrators etc will rectify such faults. This paper suggests strategies for identifying common faults and increasing the chances of successful adoption of automated solutions.

A Model of Examiner Performance

Based on a survey of the most common IT and word processing examinations in the UK (OCR 1996; Pitman 2000; City-&-Guilds 2001; OCR 2001), a model of human examiner performance emerges (Long and Dowsing 2000). These schemes focus on authentic assessment where candidates perform a series of tasks using industry standard software with considerable freedom in terms of task order and choice of alternative methods. Final outcomes, or various intermediate stages of outcomes, are assessed. The term "authentic assessment" has been employed to differentiate it from the traditional question-based assessment approach, seen throughout the U.S., for example, in the form of objective tests and multiple-choice questions. It also goes beyond commonly available computer-based assessment of I.T. skills which employ functions tests, that is, atomic interactions with (often simulated) software-specific interfaces with limited functionality. The vocational qualifications delivered by OCR Examinations aim to assess 'competencies' rather than simply 'functions' and this can be defined as the application of skills within a vocational context. It is these features that lead to a vendor-neutral form of assessment based on tasks and holistic judgement of outcomes. Although clearly more valid in terms of vocational competence this approach results in a complex markscheme that can be difficult for human examiners to apply with the required objectivity and consistency. To ensure that the required quality is maintained significant resource is applied to train examiners and monitor their performance.

The role of the examiner in authentic IT skills assessment is to apply the published assessment criteria to candidate solution documents to detect, interpret and count errors, and determine a final classification for each candidate based on the final error total, known as a fault count. The determination of the final result grade is a relatively simple process. For example, for the examination schemes targeted in the study in this paper, results are either graded by error boundaries or measured through the identification of critical and non-critical errors. The core activities of the human examiner, however, are the *detection* of errors and their *classification* so that they can be counted correctly*.* Correctly carrying out the tasks in an IT examination tend to yield a fairly well defined model solution, although some variations may be acceptable. Thus, error detection involves finding parts of the candidate document which vary from the predefined model document. Such variations constitute potential errors, though they may, in fact, relate to valid alternatives, or unimportant variations. Error detection may also encompass additional checks outside the candidate-model comparison, for example, of the consistency of use of valid alternatives within a candidate document. The classification process involves the determination of the type and context of a potential error, and its interpretation according to published assessment criteria, in order that it can be grouped and counted appropriately.

This model is presented in Figure 1. It shows the 2 principal components of the activity as error detection and error classification. Other boxes show the roles of assessment evidence and resources with respect to the assessment activity.

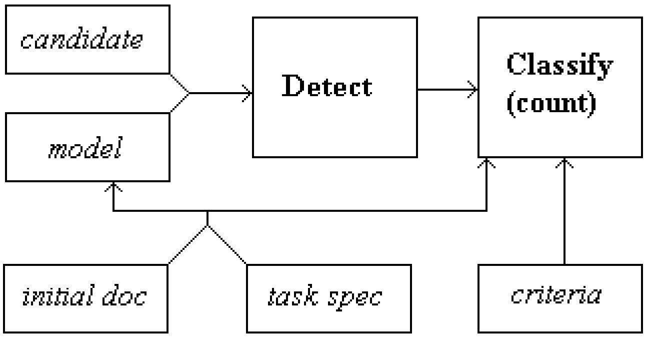

Figure 1 Human examiner performance model

Human examiners may choose to use an explicit, pre-prepared model document as reference during assessment. Many others will simply use the examination paper, which describes the tasks to be carried out, and the candidate's solution to dynamically construct, in their own minds, the relevant component of the model solution against which to judge the candidate solution. All will refer to a predefined set of marking criteria against which to classify errors once they have been identified.

Predicting Examiner Error from the Performance Model

In order to predict how human examiners may err, based on the model of performance presented above, the two principal components of performance, error detection and error classification, are considered in turn. A characterisation of the cognitive processes underpinning each is made and relevant findings from the literature considered in order to determine where the greatest risk of error lies.

Error Detection

Error detection as "Vigilant Inspection"

The task of error detection requires human examiners to apply sustained attention to study large numbers of largely similar candidate solution documents to detect relatively small and infrequent variations from the expected model document. Much research has been done in the field of Experimental Psychology to investigate human error in tasks requiring vigilance and sustained attention of this kind. Typical tasks studied include the detection of a blip on a usually clear radar screen, a warning signal in a nuclear power station monitoring system, or a faulty product on a production line of generally good quality products. Close similarities can be drawn with the nature of the detection component of the human examiner task.

In particular, error detection in IT assessment has many similarities to product inspection, which has been defined as “Vigilant Inspection”. Typically vigilant inspection tasks involve products passing along a factory conveyor belt, and hence are characterised as a search within a time limit. An alternative characterisation is that of “static inspection” where the products do not move, but are in one place in batches and work is unpaced, although inspectors must keep up with a reasonable production rate. The latter characterisation more closely resembles the examination marking task.

When inspectors are required to detect more than one signal type, “the perceptual process now becomes one of scanning, or searching, for multiple defects occurring in multiple locations...”. This closely resembles the role of the examiner when marking a solution script for errors of different kinds, for example, textual, formatting or layout errors.

Thus it is asserted that the error detection task carried out by human examiners for IT assessment can be considered a class of vigilant inspection.

Factors in Error Pattern Predictions

A ubiquitous finding of sustained attention research is the apparent drop of performance quality over time, known as the “vigilance decrement” (Davies and Parasuraman 1982) or “decrement function” (Dember and Warm 1979). A common observation is that ability to detect the signal decreases soon after performance begins and reaches its full decrement after 25-30 minutes. Clearly these findings have implications for the reliability of human IT examiners/markers, who are often required to assess large numbers of examination solutions requiring sustained attention for periods of several hours at once. An equally common finding, although one which is less well known, is the generally poor performance of people involved in attentional tasks (Warm 1984). This, perhaps, has even more alarming implications for human IT examiners/markers as it suggests that they are likely to fail to detect potential errors in candidate scripts from the outset of the assessment task, and irrespective of how many scripts they mark or how often they rest.

More particularly, work has been carried out to investigate the most common modes of error in tasks requiring sustained attention. Jacobson (1952) studied the work of 39 industrial operators who each inspected 1000 solder connections where 20 defects had been inserted (0.2% signal rate). Approximately 83% of errors were detected, so 17% were missed (type 2) while only a trivial number of incorrect identification errors (type 1) were made. The implication is therefore that IT examiners/markers are also more likely to commit type 2 rather than type 1 errors, that is, they are more likely to miss candidate errors, rather than falsely detect errors.

Empirical data is available concerning the performance of humans in tasks which are similar to the error detection task carried out by human examiners. For example, proof reading, in which text is checked for spelling and punctuation errors, clearly shares many characteristics with word processing assessment, as modelled here. Proof reading is a profession in its own right and a considerable amount of empirical work has been done on error in the proof reading task. Most work in the area has concentrated on the detection of spelling accuracy and the difference between detection of *word* errors - where the incorrect word actually forms another word, and *nonword* errors - where the misspelled word does not form a recognisable word. Professional proofreaders tend to catch around 90% of nonword errors and only around 75% of word errors. For other error detection tasks the error detection rate approaches 90% for simple mechanical errors such as mistyping a number, while it falls below 50% for the detection of complex logic errors.

Furthermore, empirical evidence would suggest that, contrary to popular belief there is not an enormous difference between the error rates of novices and experts for many tasks, including ones which require error detection.

Error Detection Reliability Prediction

Given that the error detection task performed by human markers of IT examinations requires sustained attention to detect relatively infrequent errors in large numbers of largely similar documents, it seems likely that a significant number of potential errors will be missed. The overall effect of this, if true, would be to make them over lenient. This claim is supported by empirical studies of similar tasks which show that error detection rates rarely exceed 90%, even for simple error types, and it would appear that experts, as well as novices, are prone to such shortcomings.

Error Classification

Once potential errors have been detected, where the candidate solution document varies from the correct model document, the examiner must classify them according to assessment criteria rules in order to count them appropriately. The following are fragments from a target word processing examination scheme assessment criteria:

1. Typing/spelling/punctuation errors [within words]:
One error shall be counted for each word which
1.1 contains a character which is incorrect (including upper case characters within a word),
1.2 has omitted or additional characters or spaces,
1.4 has no space following it.
…
4. Presentation
One error per examination shall be counted for occurrences of:
4A. incorrect left hand and/or top margins or ragged left hand margin,
4B No clear line space before/after separate items
4C failure to use line spacing as instructed

Assessment criteria such as those listed above can be represented as rules which prescribe how to apply error classifying and counting procedures given pertaining patterns and circumstances of differences between candidate and model solutions. For example,

> "if line spacing is not used as instructed, count a 4C error (maximum one per examination)."

The role of the examiner, having detected a potential error, is to match relevant features of the situation to the pattern components of the rules in order to activate the appropriate one.

Rules and the Mechanics of Cognition

Recent models of human cognition have used the rule analogy to describe human cognitive processes. This has emerged from the information processing model of cognition and is common throughout modern Cognitive Psychology. It is also still the dominant approach in Artificial Intelligence and the development of Expert Systems. Rules of cognition have two principal components:

1. the IF part which describes the circumstances under which the particular rule applies (also known as the "condition" or "pattern").
2. the THEN part describing what fact applies or which action to take if the rule is activated (also known as the "action").

The if/then rule structure with activation through pattern matching is the most common building block of cognition employed by Cognitive Psychologists, for example, Anderson's ACT theories have used the rule formalism as the basic structure for describing human cognition. In addition to the simple rules described above, the action component of the rule may be comprised of a set of facts or actions. The term Schemata, sometimes known as Frames or Scripts, has been widely described in the literature. Schemata are comprised of organised collections of information as well as actions which are applicable to certain situations. They can be stored and activated in the same way as the rules described above. Each schema is like a mini expert in a very specific area. Human beings employ a vast number of them, with many in operation at any given time.

The Organisation Of Cognition

Since the 1970s Cognitive Psychologists have been moving towards a unified view of, at least, the broad brush picture of the architecture of human cognition. describes this as the “emerging model of cognition”. In this model there are two components of cognition, each interacting with the other and the environment. These components are now described.

Automatic Subsystem

The automated subsystem allows a great deal of parallel access to the enormous number of schemata stored in memory. In fact, there are no known limits to the number of schemata available to the automatic subsystem. This system operates by matching the circumstances of the environment with the activation requirements or pattern components of the schemata, in the same way that rules are activated in the rule-based architecture. This processing is very fast and takes place just below the level of consciousness.

Attentional Subsystem

The attentional sub-system relates to what might be loosely considered consciousness. It has powerful logical capabilities, but at a cost. It is characterised as “limited, sequential, slow, effortful and difficult to sustain for more than brief periods”. It is now believed that the attentional subsystem also operates through schemata, and these are schemata from the automatic level which have risen to the attentional level.

Expert-Novice Differences

It is generally agreed that the movement from novice to expert involves the movement of control of cognitive processes from the attentional subsystem to the automatic subsystem. This process has been called knowledge compilation and involves mechanisms such as composition and proceduralization by which declarative, multi-step inefficient knowledge, can become single step procedural knowledge through practice and experience. In change management theory this can also be defined as the journey from unconscious ignorance (I don’t know what I don’t know), through conscious ignorance (I know what I don’t know), then conscious competence (I know but I have to think about it) to unconscious competence (I can do it without thinking). Significant time and cost is devoted to training expert examiners/markers to reach a state of ‘ unconscious competence’.

Thus an expert has large amounts of knowledge procedurally encoded so that it is almost effortlessly available. In some circumstances a novice may be able to make the same decision as the expert, but this would be arrived at via a more tortuous process of consciously trawling their knowledge. (Neves and Anderson 1981) also describe optimisations performed by experts such as memorising postulates in a geometry task, and thus avoiding search. This equates to human examiners memorising assessment criteria, and not having to refer to external resources such as marking criteria.

In summary, expertise in areas of high level knowledge is typified by knowledge compilation, that is, the encoding of knowledge in efficient rules and schemata, accessible to the automatic sub-system. This includes the encoding, through experience, of many domain and context-specific rules which allow experts to deal speedily and appropriately with a wide variety of situations.

Projecting Classification Reliability

It has been argued that the cognitive processes underpinning the error classification component of the assessment process are based on rule and schemata activation within the automatic-attentional architecture of cognition. Within this framework experts, such as professional IT examiners/markers, develop comprehensive rule sets and mature schemata which are readily available to their automatic subsystems. That is, they have instant access, without conscious search, to appropriate knowledge for a wide range of specific circumstances, and this knowledge is not prone to the resource limitation associated with the attentional subsystem. Through experience they have also developed context specific knowledge to deal with a wide range of situations. In such high level cognitive tasks, such as error classification, experts have been shown to exhibit much improved performance compared to novices or less experienced people. As discussed previously, this is in contrast to tasks requiring sustained attention, such as error detection, where experts only improve marginally on novice level performance, as they are all prone to essentially the same attentional weaknesses. In conclusion, it seems likely that the error classification process, for professional examiners, should prove much less error prone than the error detection phase. Thus a potential error, once detected, is very likely to be interpreted and counted correctly by an expert examiner.

This realisation has been used to advantage in the implementation of a semi-automated system to support the processing of the IT examinations for OCR Examinations. A sophisticated rule-based workflow solution allows the automated marker to perform all error detection activities and then categorises error classification to identify those areas that might require expert human marker intervention. The resulting system has seen improvements in efficiency (automated marking system marks scripts in seconds), initial accuracy (the first assessment is invariably correct) and consistency (all judgements will apply the same marking criteria), all of which lead to a reduced need for post-assessment standardisation and remedial action following appeal.

Process Mapping and Definition

Having established the cognitive model on which to base the expert system the first phase of the project focused on the development of the assessment engines. Once the accuracy and reliability of the assessment engines had been established the team turned their attention to the task of integrating the tool within existing processes. Migration of the entire system to a fully automated solution was considered too risky, as many Centres were not yet ready to adopt new technologies in every area of their operation. The team also recognised that adoption of fully automated solutions would require significant process and skills redevelopment within OCR Examinations and change on such a large scale was not feasible in the short-term.

A process map for the examination procedure was defined in Figure 2 and this shows the basic 'building blocks' of the system. It was important to break the process down into modules to allow a phased development and implementation programme to be adopted. Experience has shown that attempts to fully automate an entire process are seldom successful and often fall short of expectations. This can be for a number of reasons but most typically is because:

a) a particular step/module does not lend itself to automation and so the process is forced into an inappropriate solution.

b) a particular step encounters greater resistance to change due to internal and/or external factors

c) a particular step will take longer, cost more or require greater resource than the project will currently allow.

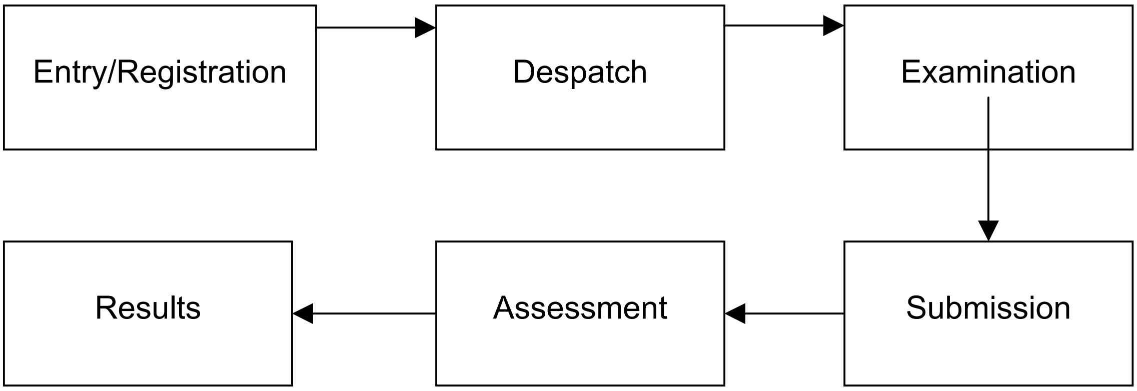

Figure 2: The EP cycle - Examination process definition model

The model identified 6 major components to the examination process, all of which could be addressed in isolation if an agreed method of describing inputs and outputs was defined. This examination process was termed the EP cycle. Underpinning all of this structure is a workflow solution that ensures that data (inputs and outputs) can progress to the next stage of the process. A series of business rules were defined that effectively created the schemata by which the complete solution could be defined. Data is progressed through a series of states where validation checks are performed to ensure that the input is as close to a notional 'goal' as possible and thereby avoid processing of incorrectly submitted data. Experience has shown that submission of information against defined instructions is prone to administrative errors but where humans are involved in the process such minor differences are corrected as the submission progresses through the EP cycle. With the introduction of an automated solution it was vital that original submissions from Centres/departments were error free or the data would be stopped at every validation stage and held up until a human referral step had been concluded. A number of utilities were developed to aid Centres/departments with the preparation of submissions and these have seen the error ratio fall from over 50% to between 10 to 15%

Entry/Registration

Two routes are available to Centres – one offers the traditional entry route using a conventional paper-based entry form and the other makes use of the secure extranet facility OCR Interchange to allow Centres to make their entries online. Both routes ensure that a Centre wishing to enter candidates for the computer based assessment (CBA) route are not sent unnecessary paperwork as the candidates' solutions will be submitted as digital files and not as printed work. These entries, whether by paper form or online, can be described as the inputs. The final trigger for despatch (the output) is the same regardless of whether the original input was paper-based or online.

Despatch

Some of the IT examinations require Centres to prepare pre-defined files for candidates to work from. These have traditionally been provided on disk or as hard copy for the Centre to generate the required files prior to the examination session. These files are now supplied via the secure extranet site for Centres to download. Question papers are still despatched in the normal way. This is primarily because there are still issues of security and restriction of exposure to the test material, which are not easily controlled through an electronic distribution system. The despatch of printed material also allows final quality assurance checks to be performed at the Awarding Body before documents are distributed to Centres.

Examination

The assessment itself is carried out on a computer using a recognised IT application. As this is already a requirement of the examination no additional conditions are being introduced to the method of assessment. Concerns over method effect when using computers to deliver conventional tests are well documented but are not an issue within the chosen suite of qualifications used for this project. Candidates are however given more precise instructions on the naming of files to ensure that data is not overwritten during collation and submission. The only additional activity required as a result of automating the system is the completion of a candidate details form (CDF) utility. This creates a file that must be submitted with the work, taking the place of the answer book, and is used to match solution files with candidate information. The creation of these files does not form part of the assessment session and can be completed outside the time allowed for the examinations.

Submission

All CBA solution files and candidate information files are submitted to OCR Examinations using the secure extranet facility OCR Interchange. This process takes a matter of seconds and the system provides an ongoing audit trail to identify the various stages of the assessment process – upload, validation, marking and process complete. Another utility, the CLAIT zipper, has been created to ensure that some first line validation takes place at the Centre and helps to reduce the number of submissions containing administrative errors. The submission procedure is not lengthy, only requires a standard Internet connection within Centres and utilises skills that will have been developed through general navigational use of the Internet.

Assessment

The IT examinations are marked automatically using the relevant assessment engines. The CDF identifies which assessment engines should be used and drop-down selection menus ensure that Centres chances of entering incorrect data at this stage are reduced. The assessment engines mark the work in seconds and where relevant any scripts requiring referral to expert human markers are flagged at this stage. Centres can typically expect results to be processed within a period of 24 hours but at various times of the day the system has been shown to process a submission and return a results report within minutes. This has seen a dramatic improvement in the service and support offered for OCR IT qualifications and Centres are further impressed by the lack of paperwork.

Results

When the process is complete a results report is generated and returned to the Centre in the conventional manner, via post for most schemes, or online to a predetermined email contact. We are currently exploring the possibility of supplying the results data via the secure extranet facility. The system has already realised a 3 week reduction in turnaround of results for some examinations.

Training and related issues

In addition to the modular approach to system development and implementation we have identified several other key factors which should be considered to ensure the success of an automated assessment solution. These have been summarised below:

a. credibility of the assessment is paramount so it is essential that the majority of the early development work is devoted to checking the reliability and validity of the assessment engines
b. identify the modules in the process that lend themselves to automation and target those first, do not try and force other modules into automation if it is not appropriate provide thorough documentation for Centres and/or departments and back this up with ‘hands on’ briefing sessions
c. establish the critical data points (inputs) in the process and try to ensure that Centres/departments are not given total freedom over entry of data in these areas ensure that internal support staff are fully trained on the system and those technical areas they are expected to support
d. consult with users during all stages of the development process.

Conclusions and future work

This paper has presented an empirical study of human examiner performance for authentic professional IT awards. A model of an automated assessment solution has been defined and strategies offered to increase the chances of successful implementation. The results of the study are now discussed and conclusions drawn. Finally, directions for future work are considered.

In more detailed analyses of modes and types of error in human examiners it has been found that errors committed by human examiners are most likely to be of Type 2, errors of omission, rather than Type 1 (commission). Analysis of performance type distributions shows that, furthermore, human assessors are most prone to missing basic candidate errors relating to the textual accuracy of solution documents.

The observed performance levels of human examiners were predicted from the model of human examiner performance presented earlier. The detection component of the human examiner task requires sustained attention as large numbers of very similar examination solution documents are studied in order to identify relatively infrequent occurrences of error. It has been shown that human beings are quite poor at such tasks and that their performance worsens over time. Theoretical models of attention provide explanations for these failings in terms of limited resources of the attentional system and the effect of the level of arousal, which can be reduced by both internal and external factors. In addition, and contrary to popular belief, there is not an enormous difference between the error rates of novices and experts for many tasks which require error detection. While textual accuracy might be considered to be the most basic and fundamental component of word processing skill, it is also one of the most difficult areas in which to detect errors for human examiners. This is because textual errors tend to stand out from the page less clearly than errors in format and layout. They require close reading of the text with full concentration, this in the context of the large work load which examiners often endure. Humans are generally good, however, at formulating a fast approximate overview of the structure of a visual object. Layout and format errors, therefore, lend themselves to human detection more readily than most text based errors.

Once potential errors have been detected, however, it has been found that examiners are extremely unlikely to interpret and count them incorrectly. This can be explained through modern cognitive science theories of learning and expertise. It is generally agreed that the movement from novice to expert involves the movement of control of cognitive processes from the attentional subsystem, which is slow and inefficient, to the automatic subsystem through which an almost limitless number of schemata can be accessed quickly and accurately. Expert professional IT examiners/markers develop comprehensive rule sets and mature schemata which are readily available to their automatic subsystems. That is, they have instant access, without conscious search, to appropriate knowledge for a wide range of specific circumstances, and this knowledge is not prone to the resource limitation associated with the attentional subsystem. The main error risk at this stage is inconsistency across experienced examiners in classifying complex or unusual errors. Even so, a potential error, once detected by an experienced examiner, is likely to be interpreted and counted correctly.

A CBA system has been developed, and is described elsewhere (Dowsing, Long et al. 1996; Dowsing and Long 1999b; Dowsing and Long 1999a), to automate the assessment process described here. Its error detection component is based on comparison of candidate and model files to generate raw errors which are classified into assessment errors in a second process through the rule-based application of assessment criteria. Empirical findings for the system have shown that it is very unlikely to miss a potential error, but that it is more likely than the human examiner to misclassify an error (Long and Dowsing 2000; Long 2001). A hybrid system has therefore been developed which allows the CBA system to carry out the bulk of the assessment, but to refer to human examiner mediation when a critical error is detected which falls into a category deemed as difficult to classify (Long and Dowsing 2000; Long 2001).

It would be interesting to carry out a study under experimental conditions to investigate human error and cognitive science issues, such as the affects of workload, level of reward, and other external factors on human examiner performance. A more controlled study could also employ more foolproof methods for the determination of definitive assessment results. Predefined error scenarios could be introduced into model documents, and assessed by human examiners under controlled conditions, to provide a set of test data. Finally, it would be interesting to carry out similar studies for other assessment schemes to determine whether results are generalisable. We intend to role out similar solutions during the next few years and the concepts defined here are already being used to support the CLAIT scheme and word processing examinations, which attract over 350,000 candidates per annum. Further studies will explore the application of CBA solutions across a wider range of qualifications and consider the use of different assessment models and processes.

References

Anderson, J. R. (1981). Cognitive skills and their acquisition. Hillsdale, NJ, Erlbaum.

Anderson, J. R. (1983). The architecture of cognition, Harvard University Press.

Davies, D. R. and R. Parasuraman (1982). The psychology of vigilance. London, Academic Press.

Dember, W. N. and J. S. Warm (1979). Psychology of perception. New York, Holt, Rinehart and Winston.

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

Dowsing, R. D. and S. Long (1999b). An Evaluation of the Impact of AI Techniques on Computerised Assessment of Word Processing Skills. the 9th International Conference on Artificial Intelligence and Education (AIED 99), Le Mans, France, IOS Press.

Dowsing, R. D., S. Long and M. R. Sleep (1996). "The CATS word processing skills assessor. Active Learning." Active Learning 4.

Long, S. (2001). Computer-based Assessment of Authentic Word Processing Skills. Doctoral thesis, School of Information Systems, University of East Anglia.

Long, S. and R. D. Dowsing (2000). Building a Computer-based Assessor for IT Skills with Enough Intelligence. International Conference on Intelligent Systems and Applications (ISA 2000), Wollongong, Australia, ICSC Academic Press.

Neves, D. M. and J. R. Anderson (1981). Knowledge Compilation: Mechanisms for the Automatization of Cognitive Skills. Cognitive skills and their acquisition. J. R. Anderson. Hillsdale, NJ, Erlbaum: pp.57-84.

OCR (1996). RSA Series Examinations Handbook. Word processing 1 part 2 (A.3.a.28). Coventry.

OCR (2001). OCR Computer Literacy and Information Technology Stage 1 (CLAIT). http://www.ocr.org.uk/schemes/it/claithome.htm

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