An Analysis of the Differences Between Traditional and Computer-Based Assessment of IT Skills

Roy Dowsing and **Stewart Long**
School of Information Systems
University of East Anglia,UK
rdd@sys.uea.ac.uk

Patrick Craven
OCR Examinations,UK
craven.p@ocr.org.uk

Abstract

Assessment performed by human examiners of candidates taking IT examinations is prone to error which, in some cases, has led to incorrect assessment results. One approach to reducing human error is to use computer-based assessment methods but these are also prone to error, albeit of different types, since computer-based methods have different strengths and weaknesses compared to human examiners. This paper compares the results of the computer-based assessment and human markers on sets of candidates sitting spreadsheet and database examinations as part of the pilot study for computerising a series of professional IT examinations.

Keywords

Assessment, IT skills, Spreadsheet, Database

Introduction

Several different types of examination are used to test IT skills. The simplest form of test is the multiple choice question (MCQ) (Haywood, 1989) but this does not test practical skills, only knowledge and therefore is rarely used. The simplest form of practical test is the function test (Brown et al., 1999) where a candidate is asked to invoke a specific function of the IT software being used, for example, to enter a value into a cell on the spreadsheet. Such exercises test the candidate's practical ability to use the software but at a very simple level and are frequently used to reinforce learning in formative assessment but rarely used in summative

assessment. A more comprehensive test is to ask the candidate to undertake a practical exercise such as that which would be expected in a typical work environment. This is called authentic assessment (Mager, 1990) and is the most common form of summative assessment for IT skills. An example of such a database test would be a set of instructions asking a candidate to enter information into a database, to correct any errors or make specified changes and then to generate reports of records which match given criteria.

A typical model of IT skills assessment is given in (Fletcher, 1992):

- Definition of assessment objectives
- Collection of evidence
- Matching of evidence to objectives
- Making judgements based on the matching

There are two choices of evidence for automated assessment; either collect the actions taken by the candidate in undertaking the test - the event stream (Dowsing, 2000) - or collect the final document produced by the candidate. Although it is possible to collect both the event stream and the outcome of a test, traditional assessment usually focuses on the outcome since this is the both easier to collect and to assess. Collection of the evidence involves collecting the output from the test in the form of paper for a human examiner and in the form of a saved file for computer-based assessment. The assessor has to determine the correctness of the answer by comparing the candidate's answer(s) to the correct answer(s). A human examiner will perform the comparison by eye whilst a computer-based assessor will use a form of string comparison (Gusfield, 1997), two dimensional for spreadsheets and databases. The result of this comparison, both from the human and computer-based assessor, is a list of differences between the two documents which can then be categorised using the specific assessment criteria relating to that examination (Dowsing et al., 1998).

The difficulty of the assessment depends largely on the complexity of the criteria and the number of errors the candidate makes. The reason that the identification of differences becomes more complicated as the model and candidate answers diverge is because synchronising – pattern matching - the candidate and model answer becomes more difficult. The difficulty of error categorisation depends on the complexity, incompleteness and ambiguity of the assessment criteria since this affects the 'intelligence' needed to apply the rules.

Different properties are exhibited differently by paper and file output, for example, formatting is partly represented on paper by the position of text whilst the same information is represented in a file by special formatting codes. This suggests that different criteria are required for human and computer-based assessors but this may not be desirable, especially where a mixture of human and computer-based assessors is used to assess the same examination or where computer-based assessment replaces human examiners. In order to ascertain the problems of using human examiner criteria for automated assessment an analysis of the criteria for a set of IT skills examinations was performed, automated assessors were implemented and the results compared with those of a human examiner for a given set of candidates.

Criteria

The test chosen for our pilot study was the Oxford, Cambridge and RSA (OCR) Computer Literacy and Information Technology (CLAIT) examination (CLAIT, 1998). OCR is one of the three major Examination Boards in the UK examining candidates in a wide variety of disciplines, including vocational skills. Approximately 300,000 candidates each year sit CLAIT examinations in a range of IT skills from word processing to music technology. This paper describes an investigation into two of the most popular application areas; database and spreadsheet tests.

The CLAIT examinations in the use of databases and spreadsheets consist of an examination paper which specifies the operations which the candidate must attempt, comprising of the instructions to construct a particular spreadsheet or database, followed by instructions on how to modify the spreadsheet or database previously input. The candidate is instructed to save the spreadsheet or database at the end of the input stage and again at the end of the editing operations. Additionally, for the database examination, the candidate has to save the results of selection and sorting operations.

The assessment criteria for the spreadsheet and database exercises are given in Tables 1 and 2. The information in the tables is supplemented by sets of rules which specify how special cases are to be dealt with. The problem for the developers of the criteria is that it is very difficult, bordering on the impossible, to predict all possible errors a candidate may make whilst taking the test and thus the assessment criteria are incomplete and/or ambiguous. In traditional examinations the problem is overcome by the examiner using ‘intelligence’ to interpret the assessment criteria and

by the use of standardisation meetings. Computerised assessors need to be constructed either with in-built intelligence or they need to refer difficult cases to a human examiner. Using human beings to solve difficult assessment problems is the more practical solution. This does presuppose that only a small number of cases will be referred to human examiners.

Certification elements		**Assessment Objectives**	
2.1	Create database and enter data	2.1.1	Initialise Application
		2.1.2	Create record structure
		2.1.3	Enter data
2.2	Edit Text	2.2.1	Edit data
		2.2.2	Add a record
		2.2.3	Delete a record
2.3	Manipulate data	2.3.1	Sort records alphabetically
		2.3.2	Sort records numerically
		2.3.3	Select records using single criteria
		2.3.4	Select records using more than 1 criteria
		2.3.5	Present records from selected records
2.4	Save, print and exit application	2.4.1	Save data
		2.4.2	Print data
		2.4.3	Exit application with secure data

Table 1:Assessment Criteria for CLAIT Database Assessment

Certification elements		Assessment Objectives	
3.1	Create spreadsheet and enter data	3.1.1	Initialise Application
		3.1.2	Enter text
		3.1.3	Enter numeric data
		3.1.4	Enter formulae
3.2	Edit and manipulate spreadsheet	3.2.1	Edit spreadsheet data
		3.2.2	Replicate entries
		3.2.3	Extend spreadsheet
		3.2.4	Generate new values
3.3	Use spreadsheet display features	3.3.1	Left and right justify text
		3.3.2	Change column width
		2.3.3	Use integer and decimal formats
3.4	Save, print and exit application	2.4.1	Save spreadsheet
		2.4.2	Print spreadsheet
		2.4.3	Exit application with secure data

Table 2: Assessment Criteria for CLAIT Spreadsheet Assessment

This study attempts to answer the following questions:

- What are the quantitative differences between the human and computer-based assessor?
- Do the differences imply that the error bounds using the computer-based system need to be modified if the same criteria are used for human and computer-based assessment?

Analysis of the Results

With the exception of one Centre, the same Centres submitted work for both tests. Many, but not all, of the candidates submitted work for both subjects. One unexpected result of both the spreadsheet and database tests was that there were no cases where the automated assessor and the human assessor differed on the classification of errors, given the same synchronisation of model to candidate. This was unexpected since the results of similar tests on word processing (Long, 1999) had shown that the automated assessor and human examiner differed considerably. Cases where the human examiner and the automated assessor produced different assessment were either due to the human examiner missing some errors or because the human examiner and the automated assessor obtained different synchronisation between model and candidate.

Analysis of the Results of the Spreadsheet Test

Thirteen Examination Centres supplied entries which were both assessed by human examiners and by the automated assessor. A summary of the results obtained is shown in Tables 3, 4 and 5. There were a number of differences between the results of the human and computer-based assessors which were due to operational problems.

CENTRE NUMBER	No of Candidates	No of Errors missed by HE	No candidates with different HE/CAA assessment	Errors made by each candidate	Total no of errors
1	4	1	0	5,1,2,15	23
2	7	3	0	2,1,0,1,1,1,0	6
3	13	4	1	1,2,9,6,2,8,9,17,7,7,3,1,34	106
4	11	2	0	0,16,8,4,0,31,10,7,0,0,17	93
5	2	0	0	0,0	0
6	1	1	0	1	1
7	1	0	0	0	0
8	4	1	0	14,1,7,7	29
9	5	0	0	0,0,0,1,1	2
10	3	0	0	1,2,1	4
11	5	6	0	9,9,1,0,10	29
12	3	0	0	2,0,1	3
13	13	11	1	0,1,2,0,12,2,8,0,7,1,0,11,15*	59
Total	**72**	**28**	**2**		**355**

Table 3: Results of the comparison of human and automated spreadsheet assessors

CENTRE NUMBER	No. of 3.3 errors	Total number of 3.3 errors	No of 3.1.2/ 3.1.3 errors	Total no of 3.1.2/3.1.3 errors
1	0,0,1,7,0	8	3,1,1,0,0	5
2	0,0,0,0,0,0,0	0	1,1,0,1,1,1,0	5
3	0,0,7,3,0,8,9,7,7,7,0,0,14	62	0,2,1,3,2,0,0,3,0,0,0,1,6	18
4	0,14,8,0,0,7,7,7,0,0,14	57	0,1,0,1,0,0,3,0,0,0,3	8
5	0,0	0	0,0	0
6	0	0	1	1
7	0	0	0	0
8	7,0,7,7	21	0,1,0,0	1
9	0,0,0,0,0	0	0,0,0,0,0	0
10	0,0,0	0	0,1,0	1

11	7,0,0,0,7	14	2,1,0,0,3	6
12	0,0,0	0	2,0,0	2
13	0,0,0,0,20*,0,0,0,0,0,0,6,20*	46	0,1,1,0,1,1,2,0,1,1,0,5,0	13
Total		**208**		**60**

TABLE 4: Results of the CLAIT Spreadsheet test comparisons (continued)

CENTRE NUMBER	Other errors	Total of other errors
1	2,0,0,8,0	10
2	1,0,0,0,0,0,0	1
3	1,0,1,0,0,0,0,7,0,0,3,0,14	26
4	0,1,0,3,0,24,0,0,0,0,0	28
5	0,0	0
6	0	0
7	0	0
8	7,0,0,0	7
9	0,0,1,1	2
10	1,1,1	3
11	0,8,1,0,0	9
12	0,0,1	1
13	0,0,1,0,0,1,6,0,6,0,0,0,0	14
Total		**101**

Table 5 Results of the CLAIT Spreadsheet test comparisons (continued)

There are a number of general observations which can be made from these tables. Firstly, the more errors the candidate makes the more likely it is that the human examiner will miss some errors. Secondly, candidates from some Centres appear to have significantly different error profiles from candidates in other Centres. There appear to be two reasons for this; candidates were at different stages in their examination preparation in this pilot study and candidates from different Centres had different backgrounds, motivation and ability. For example, some candidates were university students, some schoolchildren and some part-time adult learners.

One specific observation is that, on average, each candidate made 5 errors, although the standard deviation was high. If a candidate made more than two errors they tended to make a considerable number of errors, that is, the plot of errors for all candidates is a typical 'bathtub' curve where most candidates have either a small or high number of errors. Candidates who could do the test tended to make few errors whereas those who could not do the test made many errors; there were few in-between candidates. The majority of the errors made by the candidates (58.9%) were formatting errors (3.3 errors) and these were mainly failure to use integer and decimal format correctly (3.3.3), for example, to use the specified number of decimal places to display a number. There were fewer errors of mis-typing

numbers or strings (3.1.3 errors) (12.9%) and these were all data entry errors. The remainder of the errors, 29.2%, covered all the other types. There was only a single case where a human examiner discovered an error which did not actually exist but there were many cases where the human examiner missed errors identified by the computer-based assessor. For consistency errors, there were only 2 cases where the computer-based assessor found inconsistent use of upper and lower case letters which were not identified by the human examiner.

The difference in overall assessment between the human and computer-based assessors was just 2 candidates out of 72 (2.8%) even though the human assessors missed 28 out of 355 (7.9%) of the errors. One of the reasons for this is that textual data entry errors (3.1.2) are only penalised when the number of detected errors rises above 3 and hence any number of textual data entry errors greater than 3 would result in the same loss of objective 3.1.2. In fact the two different assessments occurred because the extra 3.1.3 errors discovered by the automated assessor were numeric data entry errors which are penalised if any exist.

Further analysis of the results – not shown in the tables – showed that the computerised assessor made a number of errors, including one case where it deduced the wrong synchronisation between the model and candidate solution. In this case the candidate had made a large number of errors, including errors in the spelling of the column names. Because of this the computer-based assessor decided that the wrong column had been deleted when in fact the human examiner thought that the correct column had been deleted. The reason for the difference was that the human used more ‘intelligence’ about what type of typing errors candidates make. In fact, in spite of the difference in synchronisation, the number and type of errors detected, and thus the overall assessment - was almost identical for both assessors in this case. Other errors in the computerised assessor were due to a misunderstanding of the criteria. An example of this is objective 3.2.2 which was interpreted by the computerised assessor as checking whether the values in cells, the contents of which had been replicated, corresponded to the model answer. This would result in an incorrect assessment if the candidate has made an error in the cell which was copied since the model answer assumes that the copied value is correct. It would also result in an incorrect assessment if the candidate had used a different formula to generate the same value as the cell to be copied. In fact, what is required is a check to see whether the formula used has been replicated, that is, whether the formula used has the same structure. In this pilot study there were no instances where a different formula had been used to

generate the correct value so the current computer-based assessor did not falsely award any 3.2.2 objectives.

Analysis of the Results of the Database Test

Sixteen Centres supplied entries which were both assessed by the human examiners and by the automated assessor. A summary of the results is shown in Tables 6 and 7.

There are a number of general observations which can be made from these tables. Firstly, as before, the more errors the candidate makes the more errors the human examiner misses. The difference in performance between candidates from different Centres was not so marked here as with spreadsheets and there is no discernible error pattern between candidates within or across Centres.

For the database examination, there is a much higher percentage of the candidates who appear to have been given the wrong assessment by the human examiner (17.8%), compared to the spreadsheet tests (2.8%). The human examiners missed 22.2% of all the errors, also much larger than the spreadsheet test (7.9%). The reasons for this are not obvious until the scripts are examined in detail when it becomes apparent that different human examiners have chosen to interpret the assessment criteria differently. The
criteria state that the information to be input into the database has to be input exactly as stated on the examination paper. One of the fields in each record was an address field and a considerable number of candidates input the shortened form of the address rather than the full form, for example, ST rather than STREET.

CENTRE NUMBER	No of Candidates	No of Errors missed by HE	Candidates with different HE/CAA assessment	Total No of errors made by each candidate	Total no of errors	Total non text/ numeric errors
1	2	0*	1	8,8	16	1
2	3	0	0	1,1,5	7	3
3	6	4	1	2,5,0,3,3,1	14	4
4	9	4	2	1,1,4,1,0,5,0,4,7	23	11
5	11	13	2	2,6,2,6,9,8,12,12,3,2,2	64	18
6	2	0	0	3,2	5	0
7	12	4	2	0,0,0,3,7,0,6,1,0,5,5,5	32	4
8	1	3	1	3	3	1
9	1	1	0	1	1	0
10	1	7	1	9	9	0
11	1	1	0	6	6	1

12	5	2	0	4,2,0,1,0	7	3
13	4	2	1	4,1,5,6	16	4
14	5	9	1	7,8,1,5,6	27	3
15	1	1	0	1	1	0
16	9	4	1	2,2,2,0,1,6,2,0,2	17	2
Total	**73**	**55**	**13**		**248**	**55**

Table 6: Analysis of the results of the CLAIT Database test

CENTRE NUMBER	abbrev/ expansion penalised	No of 2.1.3 errors	Total no of 2.1.3 errors	Other text/numeric errors	Total text/ numeric errors
1	N	8,4	12	1,2	15
2	^	1,1,0	2	0,0,2	4
3	Y	2,2,0,2,1,1	8	0,2,0,0,0,0	10
4	Y	0,0,4,0,0,1,0,3,4	12	0,0,0,0,0,0,0,0,0	12
5	N	2,5,2,3,7,5,6,5,2,2,2	41	0,0,0,0,0,0,2,2,1,0,0	46
6	Y	3,2	5	0,0	5
7	Y	0,0,0,2,4,0,2,1,0,4,2,4	19	0,0,0,0,2,0,3,0,0,1,2,1	28
8	Y	2	2	0	2
9	Y	1	1	0	1
10	N	8	8	1	9
11	Y	5	5	0	5
12	Y	2,0,0,1	3	1,0,0,0,0	4
13	Y	2,0,4,3	9	1,0,1,1	12
14	N	5,6,1,4,5	21	2,1,0,0,0	24
15	N	1	1	0	1
16	Y	2,2,2,0,0,4,1,0,1	12	0,0,0,0,0,1,1,0,1	15
Total			**161**		**193**

Table 7: More analysis of the results of the CLAIT Database test

Some examiners penalised this whilst others did not. Similar problems occurred with a title field which had some records containing MR & MRS. Some examiners penalised MR&MRS whilst others did not. The computerised assessor was consistent and penalised everything which was not exactly as specified in the model answer. This highlights a problem with the human interpretation of the assessment criteria.

The candidate error rate was 3.4 errors per examination compared to the spreadsheet error rate of 5.0. This is accounted for by the candidate's poor performance in displaying formatting in spreadsheets which does not form part of the database examination. The amount of information, which the candidate has to input in the database test, is more than twice that of the

spreadsheet test and thus it is to be expected that the number of data entry errors would be proportionately higher. The relevant figures are 0.83 data entry errors per candidate for spreadsheets and 2.2 per candidate for databases which confirms this expectation. The majority of the candidate errors were data entry errors (64.9%), 12.9% were other text errors and 22.2% were non-text/numeric errors. The computerised assessor was not error free. In one case a candidate had entered a record twice into the database. When asked to delete that particular record the candidate only deleted one of the records. The human examiner noted that the candidate had deleted the record and awarded the delete (2.2.3) objective. The computer-based assessor noted that a record still existed and penalised the candidate. The computerised algorithm has now been modified to count the number of records of the type to be deleted and award the criteria if the number has decreased after the editing operation.

Conclusions

This study has highlighted the difficulty of providing accurate assessment of this type of exercise, whether marked by humans or computers. Different human examiners interpreted the assessment criteria differently and many of them missed errors when marking the scripts. A number of errors were discovered in the computer-based assessor, partly programming ‘bugs’ and partly not foreseeing all the errors candidates made.

This study set out to answer a set of questions concerned with discovering whether it is possible to automate the present traditional assessment criteria and use these same criteria in automated assessment. The differences between human examiners using the criteria and an automated assessor using the same criteria are given in the tables above. As can be seen the human is not as accurate as the computerised system even if that system makes some errors. There were cases where the human examiner was more accurate than the computer-based system but the differences in the final assessment were so small as to be within the tolerance for most examinations. The cases where there were most differences were cases where the candidate had made numerous errors and thus had failed most of the objectives. From the data above, it appears that the spreadsheet assessment could be automated using the same criteria with very little change to the results compared to the present system using human examiners. The situation is more complex for the database assessment. Here the human examiners interpreted the assessment criteria differently

and came up with differing assessments. Thus the human examiners were not consistent and it is unclear whether the rules used by the automated system are what are required. One result of this study is that the setting specification of the database test has been changed to remove the possibility of ambiguity of interpretation of the criteria and thus such problems should not occur in future.
Another conclusion from this study is that it is impossible to make an automated assessor foolproof. The number of errors, which a candidate may make, is so large that it is impossible to envisage them all and thus an automated assessor cannot be built which can correctly cope with all possible errors. The best which can be done is to pre-test an assessor with large numbers of candidate answers, as in this pilot study, which will enable the assessor to correctly assess common errors and to continually update the assessors as new problems are discovered. Human examiners are also error prone and not only when assessing unusual submissions.

From this study the overall conclusion is that the automation of the assessment of spreadsheet and database examination is both feasible and practical, within the constraints mentioned. As a result of this and further pilot studies it has been decided that the computerised assessors are suitable for these examinations and a full implementation of the assessors will be in regular use by the Examination Board by the end of the year.

References

Brown S., Race P. & Bull J. (1999). *Computer-Assisted Assessment in Higher Education*. London: Kogan Page.

CLAIT Tutor's Handbook and Syllabus, 3rd Edition, L706, OCR, Coventry, October 1998.

Dowsing, R.D., Long, S. & Sleep, M.R. (1998). Assessing Word Processing Skills by Computer, Information Service and Use, *18*, 15 - 24.

Dowsing, R.D. (2000). Assessing Word Processing Skills by Event Stream Analysis, *International Journal of Human-Computer Studies, 52*, 453-469.

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

Gusfield, D. (1997. Algorithms on Strings, Trees and Sequences. Cambridge: Cambridge University Press.

Haywood, J. (1989) *Assessment in Higher Education.* Chichester: John Wiley.

Long. S. (1999). Outcomes of the first live pilot of WP Marker, Internal Report, School of Information Systems, University of East Anglia.

Mager R.F. (1990). *Making Instructions Work*. London: Kogan Page.

Acknowledgements

We would like to thank Sara Coldicott of OCR Examination for sponsoring this work.

Copyright © 2000 R.D. Dowsing, S. Long and P. Craven

The authors assign to ASCILITE and educational non-profit institutions a non-exclusive licence to use this document for personal use and in courses of instruction provided that the article is used in full and this copyright statement is reproduced. The authors also grant a non-exclusive licence to ASCILITE to publish this document in full on the World Wide Web (prime sites and mirrors) and in printed form within the ASCILITE 2000 conference proceedings. Any other usage is prohibited without the express permission of the authors.