Loughborough University

This item was submitted to Loughborough's Research Repository by the author.
Items in Figshare are protected by copyright, with all rights reserved, unless otherwise indicated.

Evidence, assessment criteria and the difficulty of automated IT skills assessment

PLEASE CITE THE PUBLISHED VERSION

PUBLISHER

© Loughborough University

LICENCE

CC BY-NC-ND 4.0

REPOSITORY RECORD

Dowsing, R.D., and S. Long. 2019. “Evidence, Assessment Criteria and the Difficulty of Automated IT Skills Assessment”. figshare. https://hdl.handle.net/2134/1766.

Evidence, Assessment Criteria and the Difficulty of Automated IT Skills Assessment

R.D. Dowsing and S. Long
School of Information Systems
University of East Anglia
NORWICH NR4 7TJ

Email: rdd@sys.uea.ac.uk
Tel: 01603 593102
Fax: 01603 593344

Abstract

In general, automated assessment is based on collecting evidence of a candidate's performance in answering one or more questions, relating the evidence to the correct answer or answers to determine any errors and determining the assessment by relating any errors to the given assessment criteria. In IT full-skills tests the candidate undertakes a typical exercise using a particular IT tool and the evidence collected is analysed to assess what individual skills the candidate has exhibited during the test.

One of the major difficulties of automated IT skills assessment arises from the difficulty in knowing how to associate errors made by the candidate with particular skills. The difficulties can be reduced by suitable design of the test, by reducing the complexity of the assessment criteria and by the judicious use of human examiners.

This paper illustrates the connection between evidence, assessment criteria and the difficulty of assessment with examples from word processing and the use of spreadsheets.

Introduction

During the past few years the acquisition of IT skills has become a necessity for an increasing percentage of the population as computers have become ubiquitous in many activities, both work-related and leisure. This has led to an increasing number of people wanting to acquire qualifications in IT skills with a consequent rise in the number of examination candidates. This has prompted most of the Examination Boards to investigate the use of computer-based tools to aid in all aspects of the examination process from administration and management to assessment of the candidate's work. This paper discusses the differences between human and computer-based assessment, specifically with respect to the evidence collected, the assessment criteria applied and the difficulties brought about by the introduction of computer-based marking.

Assessment is the process of establishing the level of skill or knowledge of the candidate by means of the collection and analysis of evidence. The process of assessment can be divided into the four generic stages (Fletcher, 1992). The first stage is the definition of assessment objectives and requirements to define the target skills and set the criteria specifying the level of performance or knowledge required to attain those skills. The second stage is to collect evidence to show whether the candidate has demonstrated those skills. The third stage is to match the evidence to the assessment objectives which indicates which objectives have been met and which have not. The final stage is to make judgements based on the results of the previous stages. This may involve determining whether the candidate has demonstrated enough skills to be awarded a certificate.

IT skills cover a wide range of applications with the main applications at present being word processing and the use of spreadsheets. Different people need different levels of skills in the use of these applications and thus skills need to be assessed at a range of levels and abilities. For example, professional secretaries need a higher level of word processing skills than a casual user and this requires a different form and type of assessment. Skills can be measured at a number of different levels. At the lowest level an assessment can be made as to whether a candidate is able to apply one of the IT application functions, for example, to save a file using the menu command or to embolden text in a spreadsheet cell. Such tests are known as function tests and are used widely in formative assessment. They are rarely used in summative assessment since they do not imply the ability to perform a typical task with the IT tool. Higher level tests, sometimes called full-skills test, used for summative assessment, ask the candidate to perform typical everyday tasks with a particular IT tool and the result of the test is analyzed to determine whether the candidate has exhibited sufficient skill during the test for a qualification to be awarded. This is known as Authentic Assessment (Heywood, 1989; Mager, 1990). In general, the higher the level of the tests, the more flexibility the candidate is given in performing them and the more difficult the assessment becomes.

Typically, the matching of evidence to objectives can be modeled as a comparison of the candidate's answer with one or more correct answers. This comparison takes the form of a direct comparison of the answers in IT skills assessment but may take more indirect forms, such as a comparison of properties of the candidate and model answers, in other assessments, for example, essay marking (Foltz, Kintsch and Landauer, 1998) and program assessment (Foxley, Higgins and Tsinitsifas, 1998). Differences are candidate errors and are related to the skills exhibited by the candidate using the assessment criteria provided for that skills test.

Because of the increasing number of candidates submitting for IT skills assessment, examination authorities have been looking for ways to automate the assessment. Potentially, this can reduce the cost, speed up the assessment process and provide better consistency. Automating the assessment implies that the evidence for the assessment will be collected electronically. This affects the assessment in a number of ways. Firstly, the same printed output can be produced from different sets of user actions, for example, indenting a line. Secondly, some properties of the paper output may not be available in the electronic copy, for example, where soft page breaks occur. Lastly, some properties that are difficult or impossible to measure in the paper copy are easy to assess in the electronic copy, for example, the font used. Thus using the same assessment criteria for both human-marked and computer-marked examinations is problematic.

The evidence matching phase of IT skills assessment can be divided in two tasks; the detection of potential errors in the candidate's answer and the classification of the differences into one or more assessment errors. Human examiners are good at error classification but poor at error detection whereas computer-based assessors are poor at error classification but good at error detection. The import of this is that it is possible to use a judicious mixture of human and computer-based assessors to perform IT skills assessment cost effectively (Dowsing, Long and Sleep, 1998). Another consequence of this is that as computer-based assessment moves to higher levels of the skills hierarchy, human examiner involvement becomes more necessary and important.

Evidence

Evidence of the performance of full skills tests can be collected using one of three methods:

a) Collection of the output(s) from the test as, for example, printing on paper or files on a disc,
b) Collection of the sequence of operations performed by the candidate whilst undertaking the test - the event stream,
c) Both of the above.

The type of evidence required to support a particular form of assessment differs depending on the objectives. For some types of assessment, including authentic assessment, it is advantageous to collect both the output and the event stream from a test as this can allow much more accurate analysis and matching of evidence to objectives. The event stream, together with the starting document, can provide enough information for complete assessment but its analysis is difficult. It is much simpler if the output of the assessment can be correlated with event stream actions. The printed or file copy output of the examination provides less information for the assessment but its collection is relatively straightforward and hence this is the usual evidence used for the assessment. Inferring the correlation of errors with individual function application is difficult from the output only; using the event stream as extra evidence can improve this analysis. Thus neither of the two pieces of evidence is simple to use on their own and there are advantages to having both types of evidence available.

The difference between a 'professionals' use of IT tools and an 'amateur' is principally in the method used to produce the results. Skilled users use more efficient methods to generate the result. Thus the very high level assessment of the 'professional' use of IT tools is principally concerned with the method a candidate uses to obtain a given effect as well as the outcome of the examination. For example, how a candidate formats a document or how text is deleted - via selection and backspace or via multiple backspaces - are the objective elements of the assessment. Event stream collection and analysis is essential for this level of assessment.

A further consideration is the difficulty and cost of collecting the evidence. Collection of the output from an IT test is comparatively easy whereas the collection of the event stream is considerably more difficult. In the case of human examiners, it involves the examiner observing the candidate undertaking the test and recording the actions taken. This is very expensive in examiner time and is rarely used in practice. Using computer technology, collection of the event stream is comparatively simple, for example, using hooks in Windows (Microsoft, 1995), although the use of such technology requires that the candidate's machine be instrumented to collect the data and save it for later processing. This can lead to problems of security and protection.

Assessment Criteria

Assessment criteria relate evidence to the assessment objectives, which are in turn related to the attainment of skills. Assessment criteria link errors in the evidence to the lack of individual skills. In effect assessment criteria comprise a set of rules which link the presence or absence of errors in the evidence to the absence or presence of particular skills. The generation of a set of sufficient, unambiguous and correct rules for the criteria is a difficult and error-prone task. A candidate undertaking a full-skills test is given complete use of the IT tool and thus there are an almost unlimited number of actions or combinations of actions

the candidate may make. The criteria must be applicable to all possible candidate submissions and must allow an examiner using the criteria to arrive at the 'correct' assessment.

The creation of the assessment criteria is simplified in the case of human examiners since they are assumed to apply 'common sense' to cases where the criteria are inadequate. Such cases are then reviewed at a standardization meeting to ensure that all examiners apply 'common sense' in the same way, although this does not guarantee that all errors are dealt with consistently since the standardization is only performed on a sample of the candidates (Valentine, 1932). Computer-based assessors do not possess 'common sense' and it is difficult to build in such abilities using artificial intelligence techniques. Thus where computer-based assessors are used the assessment criteria need to be specified in greater detail than for human examiners. Elaborating the criteria requires a process of knowledge elicitation from one or more human examiners (Dowsing and Long, 1999). In our experience this almost always means augmenting and clarifying the rules - intended for human examiners - to make them unambiguous and devoid of the need for human intelligence in understanding and interpreting them. As an example, a rule in a national IT examination states that a row or column in the spreadsheet has to be deleted to obtain that competence. This is adequate if the row or column is removed but how should this be interpreted if the row is only partially removed? This is left unstated in the assessment criteria and the human examiners are left to exercise 'reasonable judgement' which is validated at the standardization meeting. For computer-based assessment this implies the use of approximate matching techniques, leading to the problem of specification of the boundary between matching and non-matching.

The rules implied by the assessment criteria relate errors in the evidence (Reason, 1990) to individual skills. The evidence typically comprises a set of errors which a candidate has made during a full-skills test and one or more of the assessment criteria relate to each error or combination of errors. The assessor has to decide which assessment criteria apply to which errors. Human assessors find this relatively simple but providing rules to enable a computer-based assessor to perform the same task shows that the human assessor uses information not available to the computer-based assessor such as semantics and domain knowledge. Thus the computer-based assessor has to approximate human behaviour using only syntactic information. The document is divided into regions and required actions are associated with regions. Errors in a particular region are assumed to relate to the required task of that region and errors which relate to regions requiring no alterations are typically associated with default criteria, for example, input errors in text or spreadsheet cell contents. Such approximations can be erroneous but experiments have shown that the use of such rules allow computer-based assessors to approach human performance. Cases where it is particularly difficult to associate an error with one or more assessment criteria can be referred to a human examiner. This is one of the

reasons why function tests are simpler to assess; candidates are only asked to perform a single task so any errors made can be immediately associated with that task and the relevant assessment criteria.

There are several ways in which the performance of the sub-tasks can be linked to the overall assessment. Firstly, a simple overall grading can be applied which classifies candidates according the number and type of errors which they have made. An example, used in the RSA Word Processing Stage 1 examination, awards a distinction if there are less than three errors in the attempt or a pass if between 3 and 7 errors. Other examinations classify students according to the competencies they exhibit. In such cases the errors and omissions which a student makes whilst taking an examination subtract from the competencies certificated. An example of a competence-based IT examination is the RSA CLAIT examination. In this examination a candidate is asked to undertake tasks, completion of which indicates competence in one or more skills. Errors indicate a loss of competence either of the skill being tested or of the skill to which the error is attributed. For example, if a candidate inserts a column in a spreadsheet but with an error in one of the values then he/she would gain the insertion competence but lose the accuracy competence.

The Difficulties in changing from Human to Computer-based Assessors

Many of the major difficulties in changing from human examiner-based assessment to computer-based assessment arise because a human examiner normally assesses output printed on paper whereas a computer-based assessor examines the contents of a file. The problems stem from the fact that many different formatting commands can produce the same printed output on paper. For example, consider the centering of a heading in a word processing examination. This could be performed using the built in centering function of a word processor or by using the appropriate number of spaces or by one of numerous other methods. All these methods would result in the same appearance of the text on the paper but the saved file contents would be different. Thus a human examiner would mark all of them as correct whereas a file comparison method would only find one of them identical to the model answer. There are several different approaches to overcoming these problems. One approach is to attempt to obtain a single canonical form from all the possible ways of generating the same effect on paper. This involves converting all the format information into a single measure which is what happens in the printing process via a page description language such as PostScript (Adobe Systems, 1999) and a Raster Image Processing (RIP) engine. For most assessment this is too complex and a simpler method is required. An alternative approach is to allow the candidate to use one of a small number of alternatives and to check in the assessment for any of these alternatives. This can work well, especially in conjunction with reconsideration of the assessment objectives. In the example given above of centering a heading the question needs to be asked of which of

all the possibilities of centering a heading are really acceptable. For the assessment of printed output it is not possible to determine how the candidate has centered a heading and hence all methods are acceptable. However, with computer-based assessment it is possible to determine which formatting method was used and hence the assessment objectives need to be revisited to determine how this new information should be used. The import of this is that assessment criteria for human examiners need to be re-examined when computer-based assessment is introduced. Computer-based assessors will often need to include multiple checks for alternatives, especially for layout and formatting objectives.

The second major problem with computer-based assessment is concerned with associating errors with objectives as discussed above. In a full-skills test the candidate exhibits many different skills in some order to produce the submitted attempt. Errors in the attempt have to be related to failure to exhibit a particular skill and this requires the examiner, human or computer-based, to model the candidate's thought patterns whilst taking the examination. Human examiners have learnt what that model might be over years of experience but a programmer has to build in a simple model into the computer-based system. Of necessity this is based on simple syntactic clues rather than the richer model used by human examiners. This is where the event stream can be useful as it is a temporal record of the candidate's actions and a simple model relates errors to the current actions being undertaken by a candidate. Since event stream collection and analysis is not the norm at present, this temporal information is not available and errors are related to positional information in the document.

Reducing the Difficulty of Computer-based Assessment

There are a number of different ways of reducing the difficulty of the assessment. Firstly, restricting the examination to a series of function tests can reduce the difficulty. By doing so, all candidate actions can be associated with a single skill goal and hence this removes the assignment difficulty. The use of function tests is not generally acceptable to the Examination Boards since it does not test the appropriate skills required in the workplace, that is, it is not Authentic Assessment. Accepting that function tests are not acceptable, one way of reducing the assessment difficulty is by tightly controlling the setting specifications for the examinations. Overlapping and interacting errors are a major cause of assessment difficulty and if their number can be reduced the assessment becomes easier. The number of overlapping errors can be reduced if candidates are asked to exhibit separate skills in different parts of the document, for example, to make changes to the formulae in a set of cells in a different part of the spreadsheet to which they have to apply formatting. In many cases the skills being tested do not have to interact so changes can be made to different parts of the document. Interacting errors can be reduced in a similar fashion by careful examination design. For example, the sum of a column in a spreadsheet is affected by the result of an action to delete a row. If a candidate

does not delete a row as instructed then the column total will also be incorrect. It is not normal practice to penalize a candidate twice for the same error and hence he/she would be penalized for the failure to delete a row but not for the resulting incorrect total for the column. Suitable examination design can reduce the number of such interacting errors.

However, it must be borne in mind that candidates have considerable freedom in undertaking the examination and sometimes make changes to a document, which bear no relation to the required changes. Hence it is impossible to set an examination which can guarantee that the candidate's answers will contain no interacting or overlapping errors. However, since it is known that the majority of candidates who sit such examinations submit work which contains few errors and is thus close to the model answer, suitable examination design can restrict the difficulty of assessment.

In addition to controlling the design of an examination paper, it is possible to test the assessor with specimen answers before releasing a paper to actual candidates to determine whether there are likely to be assessment difficulties. By taking the model answer and introducing a selection of errors it is possible to test for likely difficulties. The model answer can be modified in a number of ways, for example, by introducing random errors or in a more controlled fashion by using information about the typical errors made by previous candidates. Whilst such testing can never prove that the assessment will not be problematic, it can assure the examiner that the probability of major assessment problems is small.

Lastly, the assessment criteria can be devised so as to make the assessment less complex and difficult. As stated above, errors in a test have to be related to particular functions which the candidate was expected to perform. This is usually the most difficult part of the assessment as the only evidence available, especially for outcome testing, is the final document. The smaller the number of rules corresponding to the assessment criteria the simpler, in general the assessment will be. The smaller the number of rules implies that there are fewer different classifications of errors and hence fewer problems in associating errors with user actions.

Examples

In this section we illustrate the connection between the evidence and the assessment criteria by reference to typical examples from word processing and the use of spreadsheets.

a) Word-processing

A typical word processing objective is to demonstrate the ability to move text from one position in a document to another. An example might be the following instruction given to the candidate:

In the text given below move the phrase 'along the main road to the airport' to the following paragraph immediately before the word 'scene'.

> "A policeman was patrolling along the main road to the airport when he saw a burglar leaving the house. On seeing the policeman the burglar dropped his bag of jewelry and attempted to escape by climbing the fence.
>
> Police dogs were called to the scene to assist the policeman by tracking the burglar from his scent on the bag. "

The objective is met if the text is moved successfully in the final document or if the movement is detected by analysis of the event stream collected. A problem arises when the candidate does not perform this action or not solely this action. What happens if he/she only moves part of the required text or more text than required? What happens if the candidate moves the text but with extra text inserted in the string? In order to be able to assess the skill all such questions must be able to be answered by reference to the assessment criteria. The assessment criterion in this case is 'text is moved as specified' but this does not specify what happens in 'non-obvious' cases. In the case of human examiner assessment this is left to human judgment but computer-based assessment requires extra rules. In our work we use an approximate string matching function which decides whether or not the candidate text matches the required text. If there is a match the 'move' objective is achieved and any textual errors are assessed using another criterion referring to the accuracy required of text input. If the approximate string comparison reports that there is no match then the 'move' objective fails and all the text is regarded as spurious input and assessed as such.

This case illustrates the need for approximate matching algorithms in the assessment process especially if a computer-based assessor is to match the performance of human examiners using the same assessment criteria.

b) Spreadsheets

Spreadsheets are more difficult to assess than word processing exercises because the two dimensional nature of the sheet means that interaction of errors is almost inevitable. Consider a test where the objective of assessment is to demonstrate the skill of deleting a row or column in a spreadsheet. The assessment criteria have to specify how to count errors and, because of the two dimensional structure of the spreadsheet, this is complex. The order of applying the assessment criteria is crucial. For example, if the candidate fails to delete some or the entire row other values in the spreadsheet may be affected because of dependencies. It is normal only to penalize a candidate once for a mistake so if he/she had not deleted the row/column correctly that would be counted as an error but subsequent errors in the spreadsheet due to this error would not be counted. Thus it is important that assessment units, for example, rows or columns, are assessed in dependence order. Such an order is not necessarily

the same order as the candidate is requested to undertake tasks depending on the interaction of the separate units in the spreadsheet.

The trick to performing this assessment correctly is to assess the errors in the correct order, that is, to apply the assessment criteria in the correct order, and then to correct the individual errors as they are assessed. This means that subsequent assessment will not be affected by previous errors. Consider the spreadsheet below where the candidate has duplicated the 3rd row.

	CATS	DOGS	MICE	RABBITS
JONES	1	1		
SMITH		1		3
SMITH		1		3
BLOGGS			2	

Before assessment takes place the candidate spreadsheet - rows and columns - are synchronized with the model answer. This would show that there was an extra row in the candidate spreadsheet answer. The assessment would note the appropriate error, such as incorrect values in 5 cells, and remove the offending row. The resulting spreadsheet now corresponds in rows and columns to the model. Further assessment on the candidate spreadsheet could then be performed without the extra row interfering.

Conclusions

For the assessment of IT skills, the evidence collected and the assessment criteria must be appropriate for the target skills and performance objectives being assessed. Higher level skills require more complex evidence and more complex assessment criteria, which means that the assessment process itself becomes more complex.

Computer-based assessment of full-skills tests is complex, even when full assessment criteria are provided. A major reason for this is that the candidate is given a large degree of freedom in his/her actions and therefore can submit an answer, which bears little or no resemblance to the required answer. The assessment system has to determine what the errors are and relate them to assessment criteria and individual skills, which is a complex task. Another major problem is the interaction of errors since candidates are normally only penalized once for each error and the introduction of a single error can cause multiple errors because of interdependencies.

To aid the assessment appropriate evidence is required. Many skill tests only collect and assess the output from the test although for computer-based assessment the event stream is often available. This gives valuable extra evidence to aid full-skills assessment and can help disambiguate interacting errors.

The difficulties of computer-based assessment can be reduced by tightly controlling the setting of the examination paper to reduce the possibility of interacting errors. Additionally, the examination paper can be pre-tested by assessing model answers which have had errors deliberately introduced. If such testing indicates that the assessor is likely to have assessment problems, the paper can be rewritten.

In summary, assessment difficulty is very dependent on the nature of the skills, which are being assessed. At a lower level, it is dependent on the complexity of the criteria, on the design of the examination, and on the number of errors made by a candidate. For simple criteria, such as used in HE, assessment is comparatively simple, whilst for ‘professional’ criteria the assessment is very much more complex since method matters as well as outcome.

References

1. Adobe Systems Inc. (1999), Postscript Language Reference, Addison Wesley Longman Inc, New York.

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

3. Dowsing, R.D. and Long, S. (1999), 'An Evaluation of the Impact of AI techniques on Computerised Assessment of Word Processing Skills', Proc AI-ED99, Le Man, France, 8 pages.

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

5. Foltz, P.W. Kintsch, W. and Landauer, (1998), T.K. 'The measurement of textual coherence with Latent Semantic Analysis', Discourse Processes 25, 285-307.

6. Foxley, E., Higgins, C. and Tsinitsifas, A. (1998), 'The Ceilidh System: A General Overview', Proc. Second Annual Computer Assisted Assessment Conference, Loughborough, 140 - 145.

7. Heywood, J. (1989), Assessment in Higher Education, John Wiley, Chichester.

8. Microsoft (1995), Win32 Programmers Reference, Microsoft Press, Redmond, USA.

9. Reason, J.T. (1990), Human Error, Cambridge University Press, New York.

10. Valentine, C.W. (1932), The Reliability of Examinations, University of London Press, London.