“THERE’S NO CONFIDENCE IN MULTIPLE-CHOICE TESTING,……”

Phil Davies

“There’s no Confidence in Multiple-Choice Testing, ……”

Phil Davies

School of Computing
University of Glamorgan
Trefforest
Pontypridd
Mid Glamorgan
South Wales
CF37 1DL

pdavies@glam.ac.uk

Abstract

The purpose of this paper is to introduce a study concerning the inclusion of confidence testing within the use of computerised multiple-choice tests. The modification of the O.L.A.L. (On-Line Assessment and Learning) system (Davies, 1999), permits the rewarding of the ‘more knowledgeable’ students by including the selection of confidence prior to the showing of the answer and distracters. There has been a lack of acceptance to the use of traditional multiple-choice testing, due to the student being ‘**fed**’ rather than them actually ‘**knowing**’ the answer to a question. The inclusion of confidence testing prior to the student being provided with the answer and distracters has been included in this study, in an attempt to improve both lecturer and student acceptance of multiple-choice testing.

This paper reports on the results of the study, and also provides student feedback concerning the use of this form of testing. It highlights the effect that it has had upon the student results compared with previous uses of the standard multiple-choice testing. The importance of choosing the correct weighting of positive and negative marking with respect to confidence is also highlighted.

An additional point of note is that this form of questioning has changed the emphasis of where quality is judged with respect to the creation of a multiple-choice question. The quality of the distracters currently has a major bearing upon how ‘*easy or hard’* a question is in a traditional system. With the use of this method of confidence testing, it is initially the question’s quality that decides whether a student will go for high or low marks without the need to see the possible solutions.

Introduction

'Automated assessment suffers from two problems ….Firstly it seldom makes use of information about how confident a student is in the answer given …. Secondly, it often involves the construction of complex questions to ensure that students cannot get good marks by a combination of partial knowledge and guesswork.' (Gardner-Medwin, 1995).

The use of multiple-choice questions for objective testing is becoming more prevalent in further and higher education in the UK (Bull & McKenna, 2001). It is not always for pedagogic reasons, but often for the perceived benefits of ease of management and reduced tutor marking. Lecturer acceptance is not assured, with many staff doubting the ability of multiple-choice testing to assess higher order skills, and be a fair reflection of a student's knowledge. Many staff see multiple-choice as providing the students with the answer, it does not judge their knowledge … '*knowledge is neither a dichotomous nor a trichotomous affair, which traditional multiple choice tests seem to imply, but it is continuous in the sense that there are varying degrees of knowledge*' (Echternacht, 1972).

Each question presented to the students must be fair and unambiguous. The creation of questions is not a time consuming process. What is time consuming is the creation of the correct non-ambiguous answer, and often more importantly the creation of distracters that are equally plausible, but incorrect. The quality of a multiple-choice question, could be said to be based upon the quality of the distracters, not the quality of the question.

Negative marking of multiple-choice testing is often proposed as a method to reduce guessing, etc. '*examinees generally achieve artificially high marks due to lucky guesses*' (Bush, 2001). The selection of an appropriate marking scheme is often a point of argument, in that we need to differentiate between a student who knows the answer, admits not to knowing the answer, and those who guess because they haven't a clue.

Multiple-choice testing is often used for formative purposes, but can it be used for summative assessment? Many tutors are unwilling to take the risk of using it for summative testing for reasons ranging from computer competence to doubts concerning its ability to fully test a student's ability (McKenna, 2001; Davies, 2001).

Multiple-choice testing has been used for several years within the School of Computing at the University of Glamorgan, both for summative assessment, and also as a means of promoting learning. The results have been very positive, but there has always been reluctance with the majority of staff in the department to make use of such methods due to the concern that the students are being fed the answers rather than knowing the answer. This is supported by a member of staff's overheard comment to a student *' .. don't worry about the test, it is easy, it's multiple-choice*'. This problem echoes the title of this paper, '**There's no confidence in multiple-choice testing …** '.

It is proposed that for each question asked, a student's ability in answering may fall into a number of different categories:
e.g.

a) I know it
b) I'm not quite sure, but I think I know it
c) I'm not quite sure, now I see the answers I know it
d) Perhaps I can identify the answer by a series of deductive processes on the distracters
e) If I guess then I've got a 33% chance of getting it correct (1 out of 3 system)
f) I really haven't a clue

And perhaps worse of all
g) I really know it, oh no I've got it wrong !!

'*Misinformation is particularly dangerous because the student strongly believes that the wrong answer is correct'* (Khan et al, 2001).

Standard multiple choice testing does not really differentiate the above. Student (a) should be rewarded more than the others in getting it right. The inclusion of confidence testing, prior to the showing of the answer & distracters, has gone some way to solving this problem of differentiation.

This paper proposes the use of confidence testing as a means of assessing the student's knowledge prior to the presentation of the answer and distracters. The allocation of marks is dependent upon the degree of confidence they possess with respect to their perceived knowledge prior to seeing the possible answers.

Assessment Methods

As part of the assessment process in an undergraduate level two module in Computer Communications and Networking (50% of the subject mark), the students were required to take four supervised multiple-choice tests. These tests were undertaken in weeks three (5%), six (10%), nine(15%) and twelve(20%). During the first three tests the students were permitted two passes of the tests (only one pass of the final test), with 60% of the marks from their first pass and 40% of their second pass going forward for summative purposes (Davies, 1999). This method of assessment has provided an excellent method of aiding student learning, with the initially weaker students benefiting significantly, and also confirms the positive effect this form of assessment has upon student work rates (Mulligan, 1999).

The students were presented with a question, and had to state their confidence in being able to provide the correct answer (as shown in figure 1).

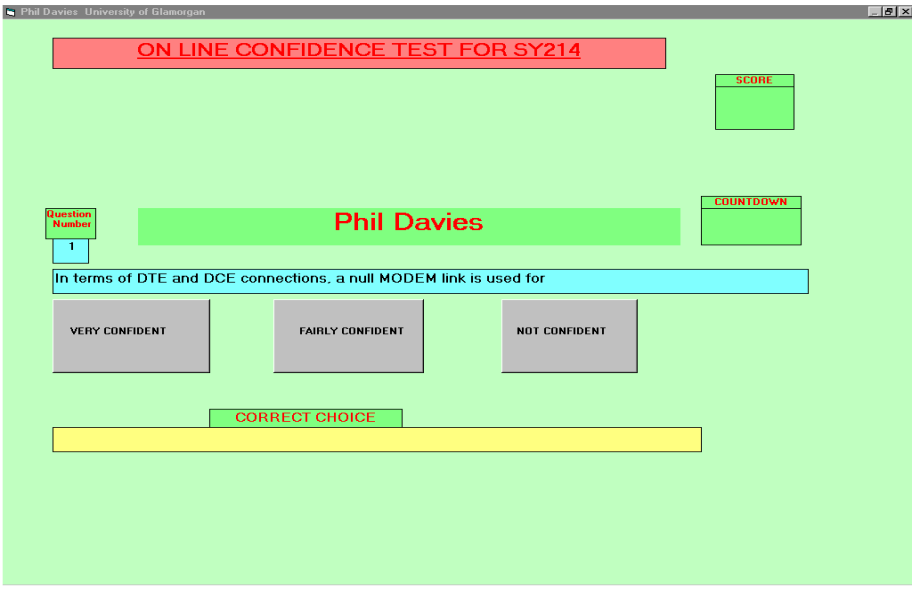

Figure 1. Confidence Test Screen Dump

Depending upon their selection of confidence, the possible marks to be awarded for each level of confidence are shown in figure 2, with included in brackets the actual associated negative marks.

	Correct	**Incorrect**
Very Confident	+4 (0)	-2 (-6)
Fairly Confident	+2 (-2)	-1 (-5)
Not Confident	+1 (-3)	0 (-4)

Figure 2. Marks Allocation per Question

Originally, +3,+2,+1,-3,-2,-1, *'equally weighted negative marking for wrong answers'* (Gardner-Medwin, 1995), were to be the marks to be used, but through discussions with the class prior to testing:

a) The students felt that those who knew the answer should receive a greater reward, hence +4.
b) If a student states that they are not sure of an answer, then they preferred 0 rather than –1 for getting it incorrect, hence 0.
c) The students were made aware that these mark weightings were provisional, and could be changed. They felt that it was important that they did not lose out in any way when compared with previous years students (addressed later in this paper).

Results

The results from the testing are given in figure 3. In order to provide a reference, the previous year's student results are provided.

Results 2000-2001 (NO CONFIDENCE TESTING)		
Test Number	**Average %**	**Standard Deviation**
OLAL TEST 1	56.49	16.11
OLAL TEST 3	49.23	13.25
OLAL TEST 4	59.18	14.25
Results 2001-2002 (WITH CONDFIDENCE TESTING)		
OLAL TEST 1	63.10	15.27
OLAL TEST 3	55.55	13.41
OLAL TEST 4 +4, +2, +1, -2, -1, 0	58.60	15.42
OLAL TEST 4 +3, +2, +1, -3, -2, 0	51.33	17.32
OLAL TEST 4 +3, +2, +1, -3, -2, -1	50.13	18.08
OLAL TEST 4 +5, +3, +1, -3, -2, -1	56.45	16.43
OLAL TEST 4 +5, +3, +1, -2, -1, 0	61.35	14.60

Figure 3. Comparison of Non-Confidence and Confidence Results (2000/2001)

From figure 3, it should be noted, that based upon OLAL tests 1 and 3, the current year's intake of students are approximately 6% better than the previous year's students (same group of questions). Therefore, it would be expected that these students' final results would be better than those of the previous year. In fact utilizing the agreed marking scheme, this year's cohort had a mean of 58.6% compared with 59.18%.

The new OLAL (with confidence testing) provides a means of identifying for each student which level of confidence was selected for each question, and whether they were correct or incorrect. By utilizing this data, different marking schemes could be applied as shown in figure 3. The results for the confidence test with the marking range of +5 to 0 at least produced an average that was better by 2% than the previous year, with a comparable standard deviation.

Figure 4 shows the percentile frequency distribution of students in the various mark groups.

Marks	2000-2001 No confidence %	2001-2002 Confidence 5,3,1,2,1,0 %
90-99	0	3
80-89	6	6
70-79	23	19
60-69	25	24
50-59	24	27
40-49	12	14
30-39	7	3
20-29	3	3
10-19	0	1
0-9	0	0

Figure 4. Percentile Frequency Distribution of Students via Mark Ranges

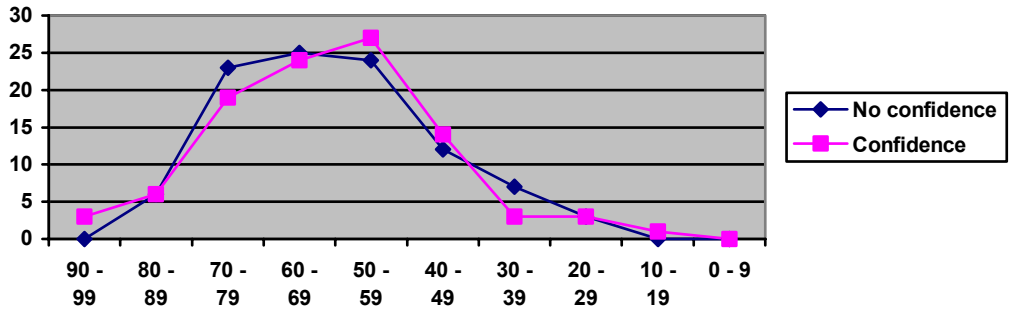

Figure 5. Percentile Frequency Distribution, Non-confidence / Confidence

It can be noted from the distributions shown in figure 5, that the use of confidence testing has increased the spread of marks, with students appearing at both the top and bottom ends of the mark scale. This is one of the proposed reasons for providing the confidence testing, to reward the brighter students and reduce the guessing factor for the weaker students. The actual graph shapes (figure 5), are fairly similar, and by selecting this marking scheme the students are mapped to the results from the previous year (i.e. did not lose out).

Comparing the results from the two years, it would be interesting to ascertain which groups of students' average marks improved from test 3 to test 4. From figure 6, the students in the range 80-89 from test 3 improved more by utilizing confidence testing. The weaker students appear to benefit less by the use of confidence testing.

Mark Range via OLAL 3	2000 (no confidence)	2001 Confidence (5,3,1,-2,-1,0)
80-89	4.3	7.00
70-79	3.85	3.59
60-69	7.32	3.63
50-59	6.37	5.07
40-49	10.52	5.80
30-39	12.44	11.10
20-29	20.71	19.55
10-19	35.05	26.11
0-9	35.39	
Quartiles		
75+	5.98	4.23
50+	6.62	3.25
25+	10.78	5.26
0+	16.29	10.49
Average Improvement	9.95	11.00

Figure 6. Average Mark Improvement, Tests 3 to 4

Having analyzed the improvement from tests 3 to 4, it is important that the introduction of confidence testing has not had a detrimental affect upon the results. It is to be expected that the student who on average answers the most questions correctly, should achieve the highest marks. Figure 7 shows the number of questions (out of the 60 questions), that were answered correctly for the students, using their marks from test 3 as a reference.

MARKS Test 3	2000 No Confidence	2001 Confidence
80-89	55.00	56.00
70-79	50.00	50.86
60-69	48.71	47.74
50-59	44.23	44.66
40-49	42.52	41.58
30-39	39.23	40.00
20-29	38.66	39.67
10-19	37.50	37.00
0-9	37.00	
Quartiles		
75+	48.48	50.58
50+	44.33	46.80
25+	42.93	43.55
0+	39.10	40.27

Figure 7. Number of Questions correct (out of 60)

This shows that the students having done well or otherwise at test 3, on average perform in a similar manner in test 4, with the students in the current year on average performing slightly better than the previous year.

Looking at the use of confidence testing, the total mapping of questions answered is given in figure 8.

	Very Confident	Fairly Confident	Not Confident
Right	59.55%	12.80%	3.16%
Wrong	11.63%	9.26%	3.60%
Total Questions	71.18%	22.06%	6.76%
Proportion Correct	83.66%	58.03%	46.75%

Figure 8. Percentile Mapping of Confidence Selection of Total Questions

It should be noted that 71.18% of the total questions answered were with high confidence, and out of these over 80% were answered correctly.

Out of the questions that were answered, figure 9 shows which groups of students based upon their test 3 results, actually showed the most confidence in selecting their answers to the 60 questions.

Test 3	Very Confident Right	Fairly Confident Right	Not Confident Right	Very Confident Wrong	Fairly Confident Wrong	Not Confident Wrong
75+	45.56	4.42	0.61	6.39	2.33	0.69
50+	36.94	8.22	1.64	7.11	4.31	1.78
25+	34.00	7.75	1.80	7.64	7.00	1.81
0+	26.42	10.33	3.58	6.78	8.58	4.36

Test 3	Very Confident	Fairly Confident	Not Confident
75+	51.95	6.75	1.3
50+	44.05	12.53	3.42
25+	41.64	14.75	3.61
0+	33.20	18.91	7.94

Figure 9. Number of Questions (out of 60) Answered via Level of Confidence

Figure 9 supports the facts that
a) the stronger students went for the very confident options
b) the weaker students went for the fairly / not confident options

Student Feedback

The students were presented with a feedback form comprising of two simple questions:

a) What are your thoughts on the use of multiple-choice testing throughout the progress of the module?
b) What are your thoughts on the use of confidence testing in the final test?

The replies to question (a) were very positive. A large proportion of the students commented on how much they had preferred the multiple-choice testing rather than having to sit an examination. A number of students commented on how the continuous testing had aided and promoted their learning in the module:

'useful learning tool'
'the second go at the test is a really useful way of learning'
'immediate feedback helps me to learn'

One of the doubts expressed previously concerning the efficiency of multiple-choice testing as a means of assessment was supported by:

'seeing possible answers jogged my memory'
'I didn't really know the answer but by a process of deduction I got the right answer'

One of the major concerns of the effect of guessing was also identified by a number of students. One student's comment summed this up,

'doesn't really reflect my lack of knowledge, I guessed and I was lucky'.

A veiled positive feedback concerning the quality of the distracters was

'It wasn't really fair because I found it difficult to tell the difference between the answers to the questions'

The only real point of negativity concerned the pressure that the students felt they were under whilst doing the test

'I felt really nervous having to answer the questions is a limited time'
'Negative marking made me panic'

The student replies to question (b), concerning the use of confidence testing, again were extremely positive. It was the first time that any of the students had been assessed in this manner, and some found it very entertaining

'I felt like a contestant on Strike it Lucky'
'I was waiting for a leggy blond to bring on my prize at the end'

A common quote that supported the use of this study was

'it eliminates guesswork'

with a number of students elaborating upon this by noting

'the system only gets the students who normally have guessed'
'it certainly reduced my guessing'

The degree of difficulty of confidence testing compared with previous testing was commented upon:

'really made me think, not like previous tests'
'very good, as it tests how much I thought I knew the work'
'really tested my knowledge, I was afraid to guess'

The stronger students appeared to appreciate the fact the they were being recognized for their ability compared with the weaker students:

'shows the shirkers from the workers'
'separated the lucky students from me'

Some of the students appeared not to appreciate the fact that by not selecting the very confident button they were automatically losing marks

'at least I didn't lose marks for getting it wrong having selected the no confidence button'

A point to note was raised by a few students, concerning what this method of assessment was testing

'if a student is not a confident person, then they will not do well'
'is it testing how confident I am, or whether I know the work'

On reflection a couple of students felt that they had approached the test in the wrong way

'I was too cautious'
'Frustrating, I realized I knew the answers when I saw them'

An important point to note was raised by a number of students concerning the type of acceptable question

'one or two questions were impossible to judge'

'need more detail and direction in the question'

One student in trying to be constructive appeared slightly confused

'perhaps it would have been better to show the answers, ask for confidence, and then show the question'

With respect to the marks allocated for each question, a number of students proposed

'the number of marks allocated for high confidence should have been increased'

Finally two comments firmly support the objectives of this trial

'I really had to think about the question, it didn't leave any space for me to guess'

'Much more like the real thing'

Conclusions

Overall, statistically and via student feedback, the use of confidence testing has been a great success. The students in general have been very positive in its use, and have felt that the marks produced have provided a fairer reflection of their abilities.

There are a number of key points that have been raised, and require further study:

a) ensuring that it is confidence in the subject area that is being assessed, rather than a person's confidence.
b) the creation of questions needs to take into account that there will be no guidance provided by the answers.
c) the allocation of marks for high and low confidence needs to be fully evaluated to ensure that the results fully differentiate the students of differing abilities.
d) whether the level of study has any affect upon the results with respect to confidence and feedback.

The feedback from the students emphasized how much they felt the use of the confidence testing had impacted upon their guessing of the answers. The general comment, especially from the brighter students, was that they now felt they were being rewarded for their ability.

The average time taken per student in answering the questions increased considerably. This matches the student comments concerning how much harder they had found the method of assessment, and how they *'**really had to think**'* when performing the confidence tests.

The combination of the basic OLAL testing, and the final use of confidence testing, has produced an integrated, fair and well-balanced assessment process, which has supported learning and fully reflected the various student abilities.

A key point identified by the students is the fact that the question is now an autonomous entity. Greater direction and specificity is required in the question than was previously the case. The initial quality of a question is based solely upon the question, not the unseen answer and distracters.

Following on from the outcome of this study, the original paper's title can be augmented to: **"There's no confidence in multiple choice testing, but it can be achieved, by utilising confidence testing within computerised multiple choice testing".**

References

Bull, J. & McKenna, C. (2001), *Blueprint for Computer-Assisted Assessment*, CAA Centre, ISBN 1-904020-00-3.

Bush, M. (2001), *A Multiple Choice Test that Rewards Partial Knowledge*, Journal of Further and Higher Education, vol 25 no 2.

Davies, P. (1999), *Learning through assessment OLAL ... On-line Assessment and Learning*, in Danson, M. and Sherratt, R. (Eds), Proceedings of the 3rd
Annual CAA Conference, Loughborough, pp 75-88.

Davies, P. (2001), *Computer Aided Assessment MUST be more than multiple-choice tests for it to be academically credible?*, in Danson, M. and Earby, C. (Eds), Proceedings of the 5th International CAA Conference, Loughborough, pp 145-150.

Echternacht, G.J. (1972), *The use of confidence testing in objective tests*, Review of Educational Research, vol 42 no 2.

Gardner-Medwin, A.R. (1995) *Confidence assessment in the teaching of basic science*, ALT-J, vol 3 no 1.

Khan, K.S., Davies, D.A. & Gupta, J.K. (2001), *Formative self-assessment using multiple true-false questions on the Internet: feedback according to confidence about correct knowledge*, Medical Teacher, vol 23 no 2.

McKenna, C. (2001), *Academic Approaches and Attitudes Towards CAA: A Qualitative Study*, in Danson, M. and Earby, C. (Eds), Proceedings of the 5th International CAA Conference, Loughborough, pp 313.

Mulligan, B. (1999), *Pilot study on the impact of frequent computerized assessment on student work rates*, in Danson, M. and Sherratt, R (Eds), Proceedings of the 3rd Annual CAA Conference, Loughborough, pp 135-147.