THE CONTINUAL ASSESSMENT OF CONFIDENCE OR KNOWLEDGE WITH HIDDEN MCQ?

Phil Davies

The Continual Assessment of Confidence or Knowledge with Hidden MCQ?

Phil Davies
School of Computing
University of Glamorgan
pdavies@glam.ac.uk

Abstract

This short paper reports on a recent study that has attempted to join together continual MCQ assessment and also the use of confidence testing utilizing hidden MCQ. Through the use of a continual assessment process of four tests of increasing magnitude with regard to the overall marks awarded, a method has been developed of identifying questions that are answered incorrectly. The questions that have been 'most incorrectly answered' have been fed into the subsequent tests. In order to identify which questions are the 'poorest' with regard to students getting them correct, a method is introduced that grades the questions making use of a confidence weighting factor.

Introduction

The iterative use of multiple choice testing has been utilized over a number of years within the module Computer Communications & Networks in the School of Computing at the University of Glamorgan, as a form of both formative and summative assessment. In a previous presentation at the CAA Conference (Davies, 1999), student results were shown to improve over the progress of the assessment process, and hence the method of testing has been identified as being of benefit to the student pedagogical process. This improvement of students via continual assessment has been noted in past proceedings (Mulligan, 1999; Sly & Rennie, 1999).

However, this method of double passing of testing using MCQ's has brought into question the academic validity of multiple choice questioning. In a subsequent CAA conference (Davies, 2002), the author introduced a method of confidence testing with hidden multiple-choice as a means of attempting to not only assess student's ability in getting a question correct, but also via a sliding scale of both positive and negative marking, assessing the degree that they 'know' the answer to a question. This use of confidence testing has been presented in various forms within other studies that have been undertaken in the past (Khan et al, 2001; Gardner-Medwin, 1995). By using this method of testing, the opportunity for guessing and attaining a 'good' score is suggested to be significantly reduced.

This paper reports upon a recent study on the integration of the iterative assessment methods (Davies, 1999) with the use of hidden MCQ (Davies, 2002), and identifies a method of grading the quality of a question by utilizing the number of correct solutions achieved and also the degree of confidence shown by the students in selecting their answers.

The confidence ratings for the three possible selections were +4, + 2, +1 for correct or -2, -1, 0 for incorrect. This relates to a student going for high confidence, getting the question wrong, and achieving a score of -6. The MCQ tests used were simple one out of three. If a student guessed at high confidence, mid-confidence or low confidence, their scores per question would be -2.66, -4.00 and -3.66 (i.e. if a student guessed all questions at high confidence, on average would score -66%). Therefore the threat of getting a question wrong and receiving a mark of -6 is a major deterrent to the student guessing at high confidence.

In order to identify whether students have improved in a particular area of the assessment process, the questions that have the 'poorest' results via the confidence rated quality grade are passed through to subsequent tests. Also a student's performance and confidence is measured throughout the progress of tests to ascertain whether there has been benefit in utilizing this iterative assessment process.

The question often associated with confidence testing is, 'are we assessing whether he/she is knowledgeable or just confident'? The comparison of the performance of student groups within the study will attempt to provide some guidance to the above question.

Methodology

Initially there were 100 students who enrolled to take this module. However, for various reasons only 88 students actually completed all four tests. The data presented in this study only reflects the students who completed all tests for comparative purposes.

Each of the tests consisted of forty multiple choice questions, comprising of a selection of one out of three. Within each test the students were expected to view a question and decide from a scoring option of +4, +2, or +1 how confident they were that they knew the correct answer. Therefore, each pass of a test being worth a maximum of 160 marks. A student having gone through one pass of a test was then permitted to re-do the test (the questions in each case being randomized in order). The maximum number of marks that were possible again for the second pass of the test being 160. However, as a method of

a) assessing a student's degree of knowing the correct answer
b) providing a deterrent against guessing the answer, a negative marking of getting the incorrect answer equated to -6, -5, and -4 depending upon the initial confidence selection of the student.

The weighting of the four tests was 10%, 15%, 20% & 25% of the module mark (the remaining 30% was from an essay type assessment). The weighting of the marks for the passes of each test was also modified as the tests progressed from 50/50, 60/40, 70/30 and finally 80/20 respectively.

In order to identify if students were improving throughout the progress of the tests, the five questions from the first test that had the 'poorest' answering were passed through to the second test. For the third test five questions were again passed through plus the five questions from the second test that were answered 'poorest'. The final test had these previously mentioned ten questions plus the five questions that were answered 'poorest' in the third test. Therefore in this way the final test was planned to have fifteen of the 'hardest' questions sat through the previous three tests, plus five questions that were based on the final few weeks of the term, and the remaining randomly selected from the previous three tests (not all exactly the same questions but in the same areas).

In order to identify the 'poorest' questions, a value was stored for each answering of a question. This value directly related to the score awarded e.g +4 to -6. Therefore, merely getting the answer to a question right or wrong was not the factor that was used, but how right or wrong using the confidence weighting.

Looking at the difference between the total weighted confidence factor of a question from the first pass to the second pass of a test may well have resulted in a very large swing. If this were the case then it was decided that this tended to indicate a question that was not of a good quality. This may have been poor wording, a distracter being too close to the correct answer, etc. Therefore, through 'trial and error' a heuristic was developed that identified on the first pass a question that had a negative weighted factor in order of magnitude (> -180), followed by a 'difference' between the first pass of the question to the second pass of the question being less than a reasonable swing factor (explained later in results section).

Results & Analysis

All of the results were assessed on the basis of 88 students who completed all four tests. The original decisions concerning which questions to carry through were not affected by the removal of the non-completing students.

On analyzing these results the following frequency distribution of student results was produced (table 1) based upon the final percentage grade achieved throughout the four tests.

Table One

80>	75-79	70-74	65-69	60-64	55-59	50-54	45-49	40-44	35-39	30-34	25-29
1	2	5	6	11	13	15	12	7	7	6	3

The overall average percentage mark for the test (excluding any other form of assessment) was 52.53%.

The overall results for the four tests are shown in table 2.

Table Two

Test #	Average Pass One	Average Pass Two	Ratio First/Second	Average Total
Test One 10%	42.56%	72.74%	50 / 50	57.65%
Test Two 15%	26.80%	66.62%	60 / 40	42.73%
Test Three 20%	41.14%	76.40%	70 / 30	51.72%
Test Four 25%	51.00%	81.50%	80 / 20	57.10%

A significant drop in the average scores achieved for test two from test one was noted. This is common with previous years of the course.

Table 3 shows via the final category grade achieved by a student, the average improvement achieved throughout the four tests making use of the weighted compensated marks achieved.

Table Three

	Test 2 - 1	Test 3 - 2	Test 4 - 3	Test 4 - 1
80>	-2	4	23	25.00
75-79	10.5	9	45.5	65.00
70-74	-31.8	43	9.6	20.80
65-69	-34.83	43.83	29.67	38.67
60-64	-45.27	53.27	16.55	24.55
55-59	-43.15	36.08	20.23	13.15
50-54	-36.2	48.8	21.53	34.13
45-49	-7.0	27.92	38.83	59.75
40-44	-45.43	30.00	28.43	13.00
35-39	-32.86	27.00	25.71	19.86
30-35	-47.33	44.5	31.33	28.50
25-29	-47.67	37.00	-13.00	-23.67
Average	**-30.25**	**+33.70**	**+23.12**	**+26.56**

Table three shows improvement from test one to test four for all but one of the above grade groups. The drop in performance from test one to test two has been noted in the past uses of these MCQ tests. It is difficult to ascertain why this drop occurs, however from student feedback they tend to find the increase in material being assessed and the expected degree of knowledge required to gain a good result in this form of assessment to be significantly more taxing than has been the case in previous modules studied. Often this has resulted in the students modifying their methods of revision, and often setting up study groups.

An analysis was then performed in order to assess the number of questions that were correct (via the 88 students) for each pass of the four tests (table 4)

Table Four

Test / Pass #	# Correct	# Wrong	Av Weighted Total using Confidence
Test 1 Pass 1	59.05	28.95	33.98
Test 1 Pass 2	74.98	13.02	202.18
Test 2 Pass 1	46.48	41.52	-73.78
Test 2 Pass 2	71.73	16.27	166.40
Test 3 Pass 1	54.48	33.52	2.70
Test 3 Pass 2	74.50	13.50	204.38
Test 4 Pass 1	61.55	26.45	73.38
Test 4 Pass 2	78.60	9.40	249.28

On examining the results from test one, the following questions were identified as having significantly negative results (table 5).

Table Five

Question Number	# Correct Pass 1	# Wrong Pass 1	Weighting Factor	# Correct Pass 2	# Wrong Pass 2	Weighting Factor	Weighting Difference Pass 2-1
Q5	27	61	-251	56	32	+9	260
Q14	30	58	-218	60	28	+67	285
Q15	34	54	-181	55	33	+4	185
Q19	8	80	-356	76	12	+223	579
Q20	33	55	-187	58	30	+16	203
Q27	11	77	-373	58	30	+21	394

However, from viewing the results of the ‘poorest’ 6 questions, question 19 (Q19) shows a significant swing from the first to second pass. This would indicate that the question had some feature associated with it that resulted in a very poor first pass, and a second pass that was better than the average. This was deemed to be a question that should not be passed through. Therefore, questions 5, 14, 15, 20 & 27 were passed through to the second test.

The results below (Table 6) indicate the 'poorest' questions from the second test through to the third test.

Table Six

Question Number	# Correct Pass 1	# Wrong Pass 1	Weighting Factor	# Correct Pass 2	# Wrong Pass 2	Weighting Factor	Weighting Difference Pass 2-1
Q2	14	74	-348	79	9	250	598
Q6	28	60	-222	45	43	-109	331
Q8	22	66	-255	70	18	145	400
Q10	28	60	-215	48	40	-66	281
Q11	20	68	-281	57	31	20	301
Q18	28	60	-200	79	9	227	427
Q19	31	57	-202	74	14	166	368
Q21	27	61	-235	57	31	32	267
Q27	32	56	-210	66	22	106	316
Q28	27	61	-256	77	11	238	494
Q32	23	65	-254	82	6	265	519

It should be noted in the above that Q27 which was previously passed through from test one is still within the range of being a poorly answered question.

The rules used to select the questions from test one to test two were: weighting in first pass >-180 and exclusion if the weighting swing is equal to or greater than 400.

In the table above, this results in questions Q6, Q10, Q11, Q19 & Q21 being passed through to test three (as well as the original questions from test one).

Table Seven

Question Number	# Correct Pass 1	# Wrong Pass 1	Weighting Factor	# Correct Pass 2	# Wrong Pass 2	Weighting Factor	Weighting Difference Pass 2-1
Q6	32	56	-212	63	25	73	285
Q11	27	61	-219	75	13	217	436
Q16	19	69	-313	60	28	58	371
Q26	30	58	-225	81	7	253	480
Q40	30	58	-218	72	16	183	401

Table 7 above shows the questions that met the rule of scoring > -180 on the weighting of pass one for passing through from test three to four. It should again be noted that two of these questions (Q6 & Q11) had previously been passed through from test two. Therefore, using the range rule of a maximum of 400 from the previous passing through of tests, no new questions would be passed through. However, as the overall results have improved in this test compared with tests one & two, the two questions Q26 & Q40 were passed through to the final test along with Q16.

Therefore, the questions to be passed through to the final test from tests one, two and three were: Q5, Q6, Q10, Q11, Q14, Q15, Q16, Q19, Q20, Q21, Q26, Q27 & Q40.

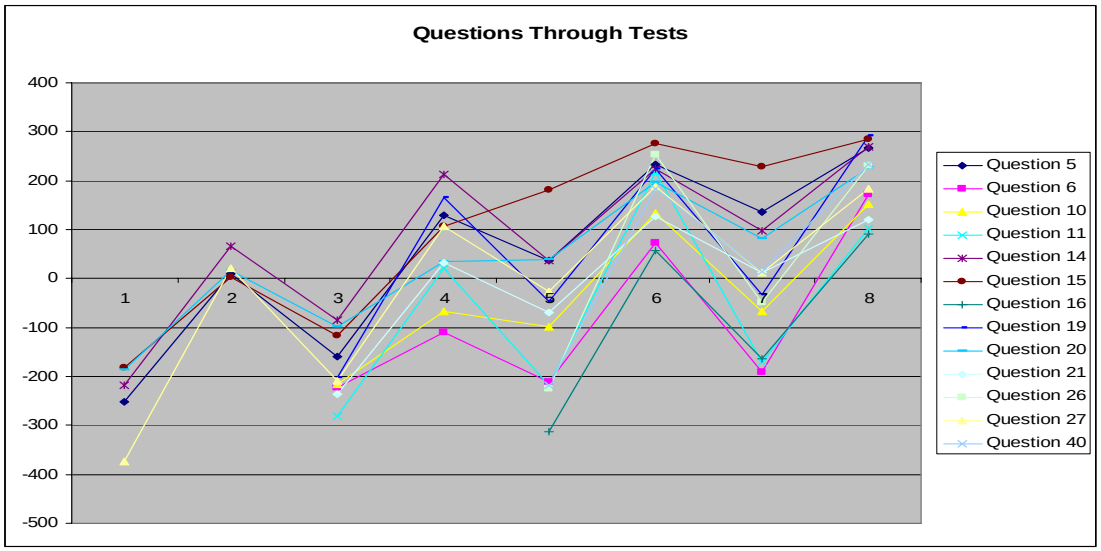
The final test therefore was made up of the thirteen ‘hardest’ questions, six completely new questions based upon the additional work that had been performed in the weeks between tests three and four, and the remainder being selected as cross-section from tests one – three.

Table 8 shows the results of the thirteen questions throughout the progress of the four tests:

Table Eight

	Test One			Test Two			Test Three			Test Four												
Question #	# Correct P1	# Wrong P2	**Weigthing**	# Correct P1	# Wrong P2	**Weighting**	# Correct P1	# Wrong P2	**Weigthing**	# Correct P1	# Wrong P2	**Weighting**	# Correct P1	# Wrong P2	**Weigthing**	# Correct P1	# Wrong P2	**Weighting**	# Correct P1	# Wrong P2	**Weigthing**	# Correct P1	# Wrong P2	**Weighting**
5	27	61	**-251**	56	32	**9**	39	49	**-160**	67	21	**130**	57	31	**37**	77	11	**232**	68	20	**136**	81	7	**266**
6							28	60	**-222**	45	43	**-109**	32	56	**-212**	63	25	**73**	33	55	**-192**	73	15	**172**
10							28	60	**-215**	48	40	**-66**	45	43	**-99**	69	19	**133**	49	39	**-67**	69	19	**152**
11							20	68	**-281**	57	31	**20**	27	61	**-219**	75	13	**217**	35	53	**-170**	64	24	**102**
14	30	58	**-218**	60	28	**67**	43	45	**-84**	74	14	**212**	56	32	**36**	75	13	**223**	62	26	**97**	80	8	**270**
15	34	54	**-181**	55	33	**4**	42	46	**-117**	66	22	**106**	76	12	**181**	82	6	**277**	79	9	**229**	82	6	**285**
16													19	69	**-313**	60	28	**58**	34	54	**-165**	63	25	**91**
19							31	57	**-202**	74	14	**166**	49	39	**-47**	77	11	**223**	52	36	**-34**	84	4	**292**
20	33	55	**-187**	58	30	**16**	43	45	**-98**	59	29	**35**	60	28	**40**	73	15	**197**	61	27	**83**	76	12	**226**
21							27	61	**-235**	57	31	**32**	49	39	**-70**	67	21	**127**	58	30	**11**	65	23	**120**
26													30	58	**-225**	81	7	**253**	49	39	**-49**	78	10	**230**
27	11	77	**-373**	58	30	**21**	32	56	**-210**	66	22	**106**	51	37	**-26**	73	15	**189**	55	33	**12**	72	16	**184**
40													30	58	**-218**	72	16	**183**	56	32	**15**	77	11	**233**

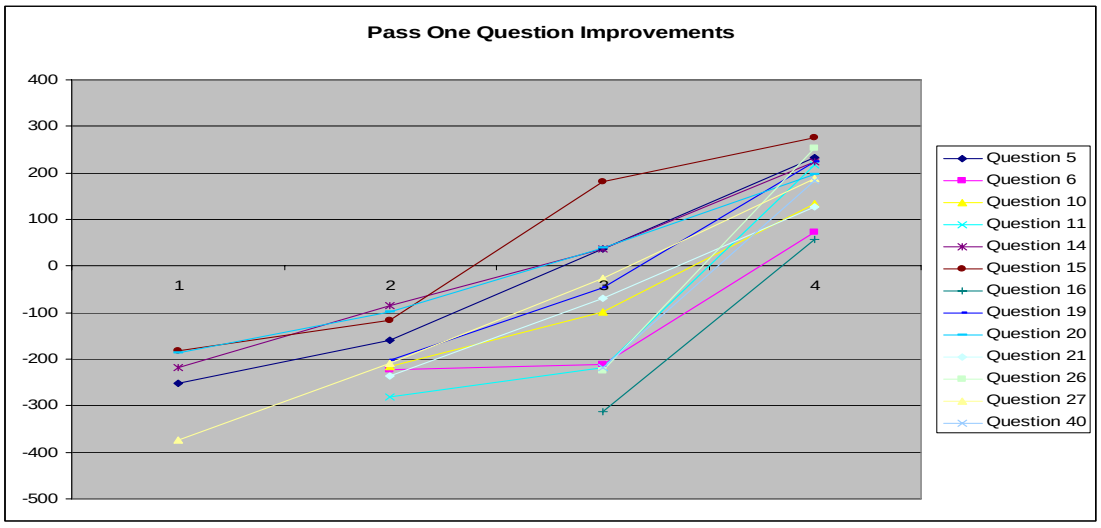
Having identified the questions, and mapped their progress throughout the series of tests, it is now important to evaluate whether the results for these questions have improved.



The above chart shows an improvement of the questions throughout the progress of the tests for the 'poorly' answered questions. This positive trend appears to indicate a learning of the material throughout the process of testing within the module.

However, it should be noted that following on from test to test, there is an initial drop in the weighting of the results. This might indicate that the improvement from pass one to pass two of a test indicates short term memory. However, if this were the only improvement achieved then the results would to be the same from test to test (which is not the case).

The chart below shows the improvement of results of the pass through questions for their first pass of results.



In order to assess whether groups of students improved through the progress of the testing, an analysis of one of the five questions (Q5) was performed. The table (table 9) is ordered via final grade category achieved by the students.

Table Nine

Q5	80>	75-79	70-74	65-69	60-64	55-59	50-54	45-49	40-44	35-39	30-34	25-29
Tst1 P1 Cor/Wr	0/1	1/1	3/2	0/6	6/5	2/11	7/8	3/9	2/5	2/5	0/6	1/2
Weight	-5.00	-0.50	0.20	-5.83	-0.91	-4.08	-2.06	-3.33	-3.00	-3.43	-4.33	-2.00
Tst1 P2 Cor/Wr	1/0	2/0	5/0	6/0	9/2	7/6	11/4	4/8	2/5	3/4	4/2	2/1
Weight	4.00	4.00.	4.00	4.00	2.18	-0.85	0.73	-2.67	-2.86	-2.29	-0.16	-0.67
Tst2 P1 Cor/Wr	1/0	1/1	5/0	3/3	6/5	5/8	6/9	5/7	1/6	5/2	0/6	1/2
Weight	4.00	-1.00	4.00	-1.33	-0.73	2.62	-2.00	-2.67	-4.29	-0.43	-5.00	-2.33
Tst2 P2 Cor/Wr	1/0	2/0	5/0	5/1	11/0	11/2	13/2	8/4	3/4	4/3	2/4	2/1
Weight	4.00	4.00	4.00	2.33	4.00	2.15	2.53	0.67	-1.43	-1.00	-3.00	0.33
Tst3 P1 Cor/Wr	1/0	1/1	5/0	5/1	9/2	10/3	8/7	7/5	2/5	7/0	0/6	2/1
Weight	4.00	-1.00	4.00	2.00	2.18	1.53	-0.40	-0.42	-3.29	3.14	-4.67	-0.33
Tst3 P2 Cor/Wr	1/0	2/0	5/0	6/0	9/2	11/2	15/0	10/2	6/1	5/2	6/0	1/2
Weight	4.0	4.0	4.0	4.0	2.18	2.46	3.73	2.42	2.57	1.14	3.00	3.00
Tst4 P1 Cor/Wr	1/0	2/0	5/0	6/0	10/1	12/1	12/3	8/4	2/5	5/2	2/4	3/0
Weight	4.0	4.0	4.0	4.0	3.09	3.08	1.67	0.50	2.86	0.43	-2.83	3.00
Tst4 P2 Cor/Wr	1/0	2/0	5/0	6/0	11/0	13/0	14/1	11/1	5/2	7/0	4/2	2/1
Weight	4.00	4.00	4.00	4.00	4.00	4.00	3.13	3.08	1.14	3.29	0.17	-0.67

Table 9 shows in the Cor/Wr rows the number of students who had the question Correct / Wrong. It is noticeable that the stronger students improved quickly throughout the tests with regard to getting this particular question correct, with high confidence. However, it should also be noted that some of the weaker students also improved, but not as quickly. In the past it has been suggested that by having continuous assessment the students would learn, and improve. This is obviously true for the stronger students, especially noted by the improvements shown not only in the second pass but in the first passes of the test. This would appear to indicate that the weaker students on average were improving from pass to pass of a test, but not as much from test to test as were the stronger students.

The weighting range could take the value from 4.00 high confidence correct, to -6.00 high confidence incorrect. The weighting figures show a significant improvement in a swing from positive to negative, indicating not only that the students were getting better, but were also showing a greater confidence in being correct.

Further analysis is currently being performed in order to quantify these improvement trends across the questions.

Table Ten

Category	Average # of Questions Correct Pass 1	Average # of Questions Correct Pass 2	Average Total # of Questions Correct
80>	84.00	92.00	176.00
75-79	78.00	88.50	166.50
70-74	81.60	96.00	177.60
65-69	77.67	91.17	168.83
60-64	75.00	88.00	163.00
55-59	73.62	87.63	161.23
50-54	73.60	88.40	162.00
45-49	70.92	84.42	162.00
40-44	69.43	80.86	150.29
35-39	66.00	80.29	146.29
30-34	59.50	78.50	138.00
25-29	62.00	76.67	136.67

Table 10 shows the average number of questions correct out of a total number of questions for each pass of 160, giving a total of 320 questions. These results do not take into account the weightings of the tests that these questions belong to, or the weighting of the passes of the tests.

On the whole these results above are very positive with regard to the mapping of the number of questions that the students had correct and their final grade achieved via the confidence ratios etc.

However, there a few students whose results do not fit in with their expected averages, and further work will attempt to identify these students, and the

reasons behind the amount of questions they had correct and their final scores. Examples of some of these are given below:

Category	Average # of Questions Correct Pass 1	Average # of Questions Correct Pass 2	Average Total # of Questions Correct
25-29	72	81	153
70-74	94	112	206
55-59	80	94	174

Conclusions

The results from this study have proved to be very positive with regard to the mapping of the number of questions a student gets correct and the weighted average score they achieve.

The method of assessment initially produced some poor student results, and hence a degree of student concern was expressed concerning the judgment of their ability via what to them was an entirely new method of assessment. However, what this indicated to a large number of students from their own feedback, was that the methods that they had used for revising in the past produced a very superficial knowledge in a subject area. The method used for this assessment had ‘really made me study harder’.

The progression throughout the tests has again been very positive and suggests that the students improve throughout the process of assessment.

The improvement between pass one and pass two of the tests has not been that great, therefore this indicates that the improvement of students is not merely short term memory, but appears to be a general improvement in performance over the series of tests.

The weighting of the questions via the confidence rating has provided a method whereby the ‘quality’ of a question can be measured to quite a fine granularity of difficulty. This now opens up the possibility of using adaptive tests in the future. However, it should be noted that if the same questions (or topic areas) are to be used in a series of tests, the actual confidence weighted value of a question may change dramatically.

Further work is being currently undertaken in attempting to identify students whose knowledge has in some cases been good, yet they have failed to achieve the expected grade associated with their ability. This can not be solely attributed to their lack of confidence in their own ability, as they may have just been lucky in getting a number of questions correct on a first pass, yet not been very confident.

Overall this study has been very successful, yet no major conclusions can be made as of yet as the analysis is at the ‘work in progress’ stage.

References

Davies, P. (1999) Learning Though Assessment: OLAL On-Line Assessment & Learning, *(eds) Danson & Sherratt, 3rd Computer Assisted Assessment Conference, Loughborough*, 75-88.

Davies, P. (2002) "There's no Confidence in Multiple-Choice Testing, …", *(ed) Danson, 6th International Computer Assisted Assessment Conference, Loughborough,* 119-130.

Gardner-Medwin, A (1995), Confidence Assessment in the teaching of basic science, *ALT-J*, 3, 1, 80-85.

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*, 23, 2, 158-163.

Sly, L & Rennie, L. (1999) Computer Managed Learning: its use in formative as well as summative assessment, *(eds) Danson & Sherratt, 3rd Computer Assisted Assessment Conference, Loughborough* 179-189.

Mulligan, B. (1999) Pilot Study on the impact of frequent computerized assessment on student work rates, *(eds) Danson & Sherratt, 3rd Computer Assisted Assessment Conference, Loughborough*, 135-147.