

On-line assessment: comparison with paper-based testing Final Report

Introduction

The project reported here builds on an earlier study (Chevins 2005: <http://www.bioscience.heacademy.ac.uk/journal/vol5/beej-5-1.htm>) in which I showed that replacing lectures with prescribed reading and frequent in course assessment improved student performance in an Animal Physiology module. In that study both formative and summative assessments were delivered on paper under examination conditions and were thus compulsory. The present project used the same teaching method and summative assessments, but presented the formative assessments on line. The aim was to compare overall student achievement, including examination performance, under these conditions with the previously regime. Because of various delays with the project, I have been able to collect three years' data using two different on-line conditions, and compare it with the previous paper-based study. As all of the teaching material was presented using a virtual learning environment (VLE) extensive information concerning student usage of the formative tests and related material is also available.

Methods

The study was carried out in a level 2 Animal Physiology module running at Keele University in the Biology degree programme in semester 1. There are nominally 12 teaching weeks from September to December, but in practice there is little teaching in week 1 or in the "reading week", week 7. The end of module examination occurs in the second or third week of January. The module content has remained constant from at least 1997-8 to 2007-08 with minor variations, and is presented under the headings shown in Table 1. All work was based on prescribed reading in Eckert Animal Physiology (5th Edition) by Randall, Buggren and French (2002).

Table 1
Schedule of Topics and in-course tests

Module Week	Topic	Formative Objective Tests	Summative Objective Tests
1	Introduction to module		
2	Osmoregulation and excretion	Formative MCQ 1	
3			Summative MCQ 1
4	Cardiovascular and respiratory physiology	Formative MCQ 2	
5			Summative MCQ 2
6	Nerve physiology	Formative MCQ 3	
7	Reading Week		
8	Nerve physiology		Summative MCQ 3
9	Muscle physiology	Formative MCQ 4	
10	Muscle physiology		Summative MCQ 4
11	Hormone physiology	Formative MCQ 5	
12	Hormone physiology		Summative MCQ 5

Prior to 2005-06 all in-course objective tests were presented on paper. In 2005-06 formative tests on the first two topics (osmoregulation/excretion and cardiovascular/respiration) remained paper-based, whilst formative tests on the latter three topics (nerve, muscle and hormones) were transferred to the VLE. As the topics are used for comparison they are referred to in future as group 1 (topics 1 and 2) and group 2 (topics 3, 4 and 5). For 2006-07 and 2007-08 formative tests for topics in group 1 were also presented through the VLE (WebCT version Vista 4) as shown in Table 2.

An important difference between presenting formative tests on paper or on the VLE is that the paper test were given on a single occasion, with correct answers, feedback and marks the next day. Presenting the tests on the VLE enabled immediate marking and feedback, and the tests remained available to revisit throughout the rest of the semester and up to examination (and later re-sit examination) time.

The summative tests remained paper-based, and the same questions were used in them throughout the project. An important feature of the whole scheme is that students knew that about 30% of the summative questions on any topic would have appeared on the preceding formative test, as an incentive to use the feedback and learn the material.

Three other important differences in the mode of presentation of the **formative** tests between group 1 and group 2 topics for 2006-07 and 07-08 need to be noted.

1. Tests in group 1 generally had less detailed feedback (explanation of why answers were right or wrong) than in group 2.
2. Group 1 test feedback immediately showed which was the correct answer (WebCT “self-test”). The feedback for group 2 tests gave no such indication (WebCT “quiz”). Students who got answers wrong at the first attempt could only find the correct answer in group 2 by trying again with or without referring to the textbook (as advised in the feedback).
3. In 2006-07 and 2007-08 the group 2 formative tests were generated by WebCT from a pool of questions about twice as large as the number of questions in any test attempt. Students thus received a different test at every attempt, with questions selected at random except that the proportion of questions of each type and in each sub-topic remained constant. They were thus encouraged to try the tests in group 2 more than once. The on-line formative tests for group 1 were the same for each attempt; there was no larger question bank.

Table 2

Year	Group 1 Topics	Group 2 Topics
2003-04	Paper-based formative tests	Paper-based formative tests
2004-05	Paper-based formative tests	Paper-based formative tests
2005-06	Paper-based formative tests	VLE-based simple self- tests
2006-07	VLE-based simple self- tests	VLE-based complex quizzes
2007-08	VLE-based simple self- tests	VLE-based complex quizzes

Principal comparisons

The main comparisons made are (i) between paper-based and simple on-line tests (i.e the same test, not involving a question-bank) and (ii) between simple and complex on-line tests (the latter involving randomised questions from a question bank at each attempt). Both types of comparison are made across year cohorts, and where possible within year cohorts.

The metrics compared are (i) the student performance in the in-course summative objective tests and (ii) in the end of module essay examinations. A preliminary analysis of pattern of student use of the formative tests on the VLE is also included.

Results and Discussion

Summative test results

Table 3

Prior formative test mode	Group 1 topic summative scores (percent)		Group 2 topic summative scores (percent)		
	Osmo/excretn.	Cardio/respn.	Nerve	Muscle	Hormone
Paper	60.3	59.4	61.7	53.6	61.3
Simple on-line	59.1 ¹	56.2 ¹	61.6 ²	47.1 ²	60.3 ²
Complex on-line	-	-	62.0 ¹	56.1 ¹	70.5 ¹

¹ Two-year mean 2006-07 and 2007-08

² Single year score 2005-06

Table 3 shows the mean scores (percent) on the summative tests for each topic following paper-based formative testing, as compared with the summative scores following simple and complex on-line formative testing. The scores show no evidence that on-line formative testing *per se* in this module resulted in improvement in scores in the subsequent summative tests. More sophisticated (“Complex on-line”) formative tests in which a different set of questions is selected at random from a question-bank had no obvious effect in one of the group 2 topics (nerve) but a produced a marked (~10%) improvement in the second and third (muscle and hormones). This suggests that students may have used the complex on-line tests and feedback to more effect in at least the latter two topics. Possible reasons for differential effects on the topics are discussed later.

Examination performance

The end of module examination scores for the period of the study have been recorded and analysed, but are unfortunately so erratic that no valid conclusion can be drawn regarding the effect of the formative test regime, so the data are not presented here. The reasons for this include differences in

ability between different student cohorts, but a variation in difficulty of examination questions is probably more important.

Although it is disappointing that examination scores yield no useful data for the main study, there are interesting differences in students' performance between the different sub-topics of the module. In particular examination essays on the nerve physiology topic result in marks with a very low mean but equally high standard deviation. For 2006-07 the class mean was 29.3+/- 24.4%. Of the 16 students who chose to answer this question 13 failed, one had a bare pass and there were two very high first class scores. In 2007-08 the results were virtually identical: class mean 29.1%+/- 22.1%. 17 students chose to answer the question 13 failed, two with scores of zero, 1 had a bare pass and there were three 2(i) answers. Students often do this topic very badly but a few do it very well. It may be that the examination questions have been more searching in this topic, demanding higher order thinking, whilst some of the other topics that proved popular student choices (see below) needed only knowledge or at most comprehension in Bloom's taxonomy. This supposition is supported by the fact that for an easier nerve physiology question set in 2005-06 the class marks were 50.3+/-24.5% with 5 fails and 5 firsts from 22 students choosing the question.

For comparison, questions on the osmoregulation/excretion topic are usually well answered with lower variance in the scores. In 2005-06 class mean 62.0+/- 14.7% (N=32); for 2006-07 mean 60.0+/-13.2% (N= 55) and in 2007-08 mean 57.5 +/- 18.0% (N=31). These questions were at lower levels in the Bloom taxonomy.

The topics on which students choose to answer exam questions is also interesting. The exam paper always contains four questions from which students choose two, so an even distribution of answers amongst the topics would result in 25% of answers on each topic. This is true within 2% for nerve, muscle and cardiovascular/respiratory physiology, but only 7% of students chose the hormone topic whilst 41% chose osmoregulation/excretion over the years 2004-2007. It is hypothesized that the exam questions on the latter topic have not only been undemanding but also rather predictable, perhaps leading to some question spotting by students. But the reason for low take-up of endocrinology questions is not clear especially as the summative in-course test results for this topic were so good.

Student study patterns

All the factual information in this section is derived from the "tracking data" available in WebCT (now Blackboard) which is a rich source of information on student learning activity. The system provides information on how many times each course item (e.g. self-test or quiz) was used (no. of visits) the average length of time of each visit and the total time spent on each item by all class members. This information can be given for individual named students, and for any period (day, week etc) so the time-course of student activity can be plotted. In addition, for the quizzes used for the more complex formative tests the system shows an additional record of individual attempts at

the quiz. This seems to be *in addition* to the data for “visits” (above) but at the time of writing this report it is not clear, and clarification is being sought from WebCT/Blackboard. For the purpose of this report it is assumed that the single data set (visits) for the simple self-tests is equivalent to the two sets of data (visits plus attempts) for the complex quizzes.

Overall use of formative objective tests throughout the module is shown in Figure 1 for 2007-08, using only the “visits” for the quizzes (excluding “attempts”). The figure shows that students use the more complex quizzes more often than the simple self-tests by a factor of 2 – 3.

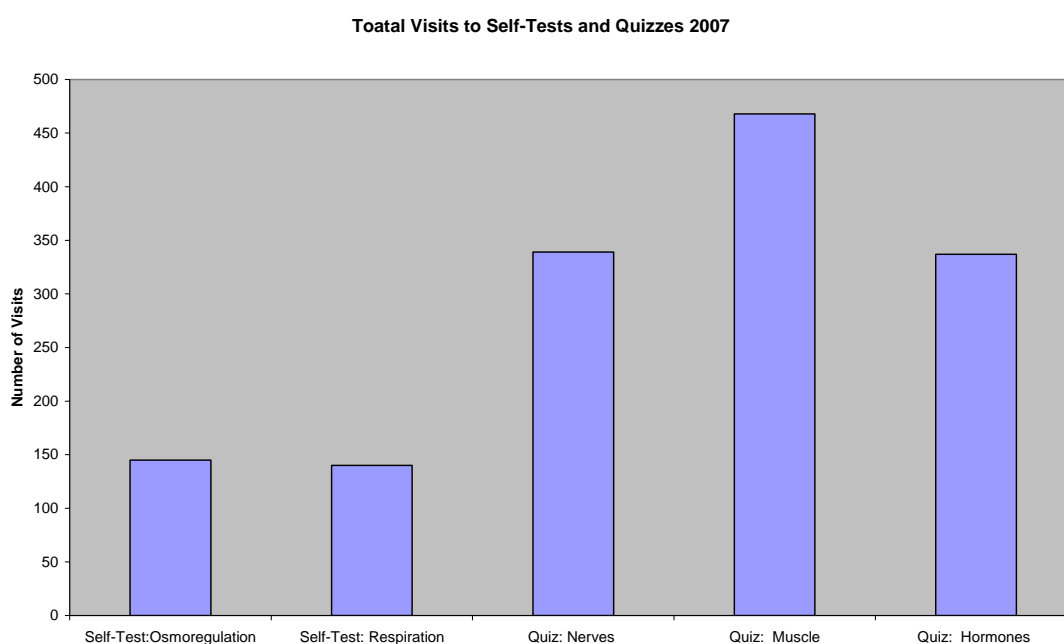


Figure 1

However, when the use of tests is expressed in terms of time spent on each (figure 2 below) it is apparent that students spent less time at each visit for the quizzes, so that the total time spent on quizzes is only marginally more than the total time spent on the simpler self-tests.

This may be an artefact of the way in which WebCT records the data. For the quizzes visits may represent only initial opening of the quiz and/or reading the results of each attempt with the feedback. The time spent actually answering the quiz is probably what is recorded as the time for each attempt. This remains to be verified by the vendors of the software, but is born out by the fact that each early attempt can take 10 minutes or more. This has usually halved by the time a student is on his/her 10th attempt, and for those that go on to make 15 - 20 attempts, the later ones are usually of only one or two minutes' duration. By this time a student is probably mainly scanning for novel questions, or checking familiar ones if there is still uncertainty about the correct answer. The records for each separate attempt shows that for 2006-07 half the students used each quiz at least 4 times and nearly 20% of them used each quiz over 10 times. The figures were similar for the 2006-07 cohort. Occasional students use each quiz 20 or more times.

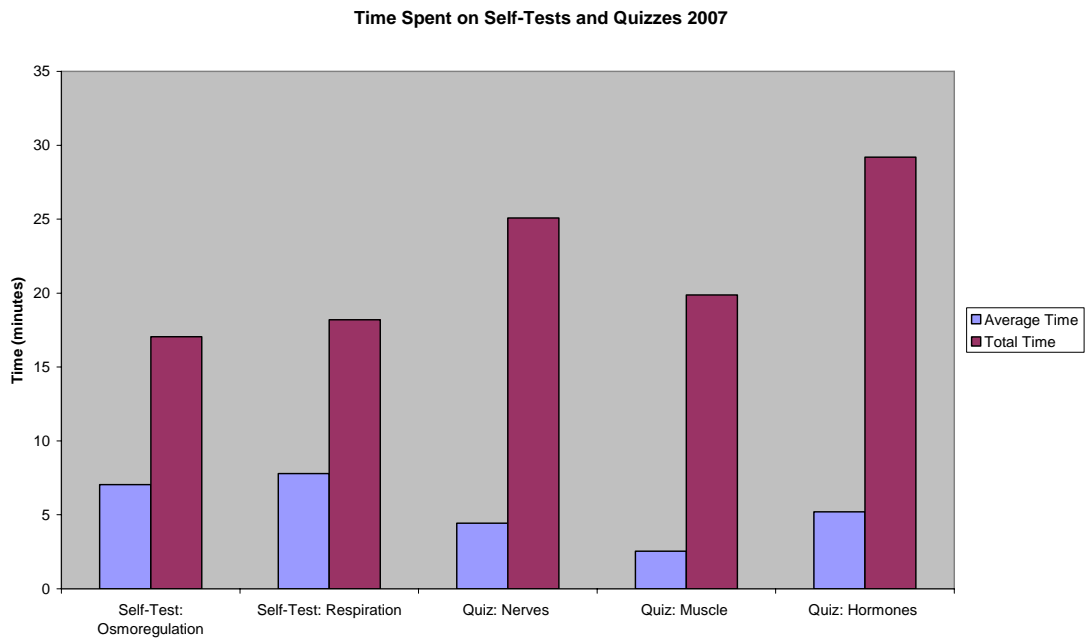


Figure 2

Figure 3 (below) shows the usage of self-test and quizzes when the number of attempts are added in. Interpreting the data in this way shows quiz usage as 4 - 5 times more frequent than the self-tests.

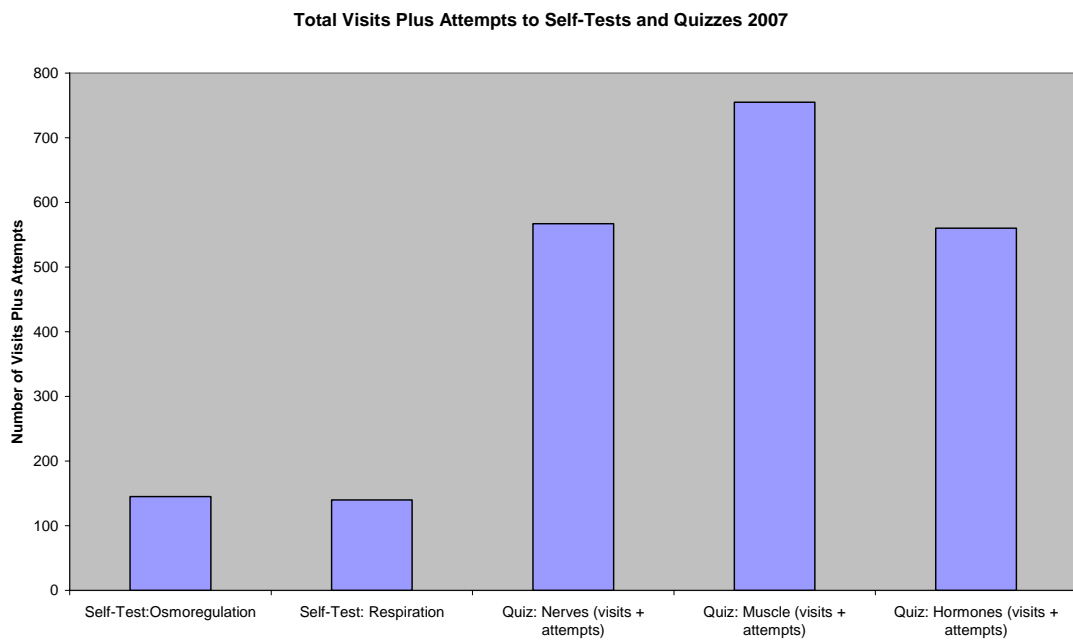


Figure 3

Pattern of Self-Test and Quiz use.

The pattern of student use of the quizzes and self tests over time has been analysed in a number of ways. First, the data extracted for use of the self-tests for the group one topics on the days leading up to the related summative is shown in Figure 4

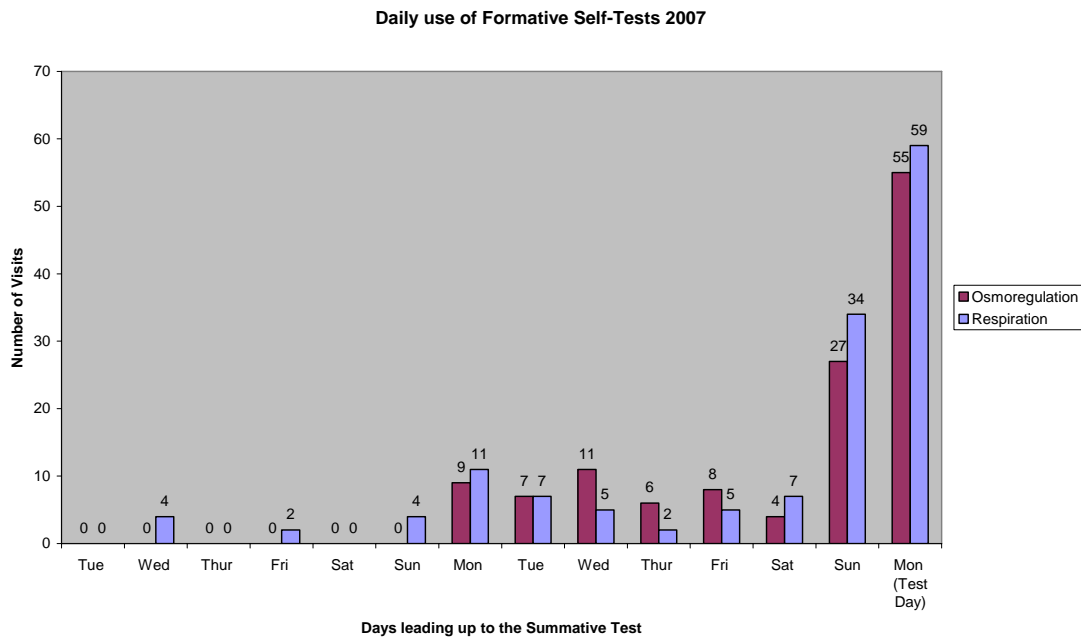


Figure 4

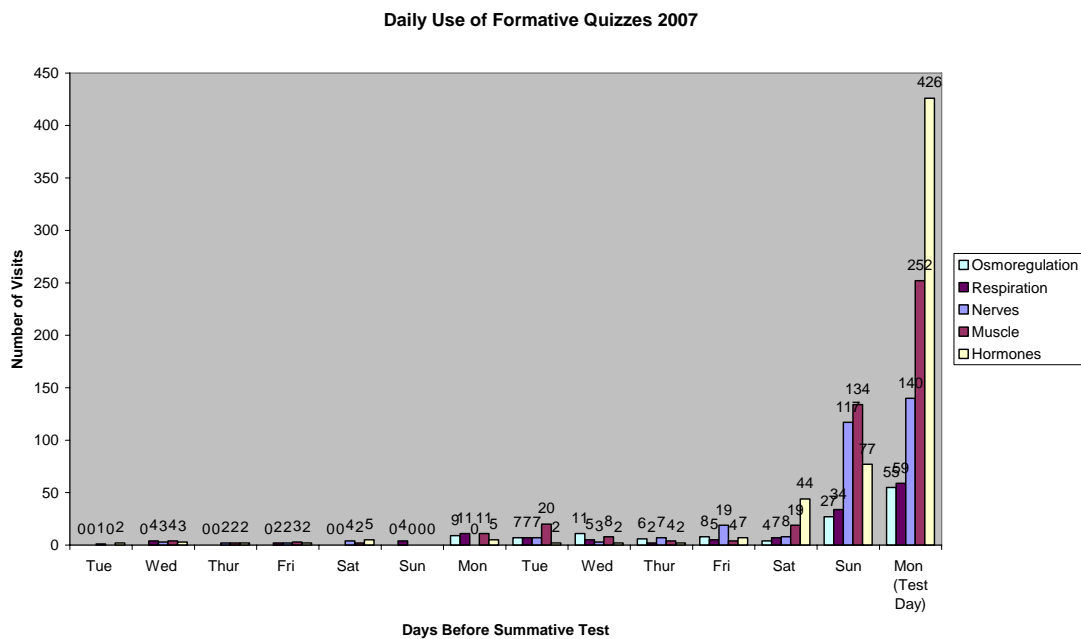


Figure 5

Figure 5 is a data for the same student cohort in 2007, but adding in the group 2 topics that used the complex quiz. Note that the scale in figure 5 is different

from figure 4. The numerical data have been shown to aid comparison between the figures. The main differences are that a few students used the quizzes earlier than the self-tests in the two week study period for each topic. The patterns are then similar until the 2 – 3 days before the summative test, at which time the use of quizzes increased much more than the self-tests.

Which ever way it is measured, it is evident that students make much more use of the quizzes for the group 2 topics than they do of the self-tests for the group 1 topics. As some of the group 2 results improved after introduction of the complex quizzes (summative test results above) it seemed important to try to discover whether this was a causal effect. One way to investigate it is to look for correlations between the test scores and time spent on formative tests for individual students. This has been done for the five topics during the seven days leading up to each summative test. Group 1 topic data is shown below as the two graphs in Figure 6. No correlation is apparent and Pearson's R analysis is not significant at the 5% level in either case.

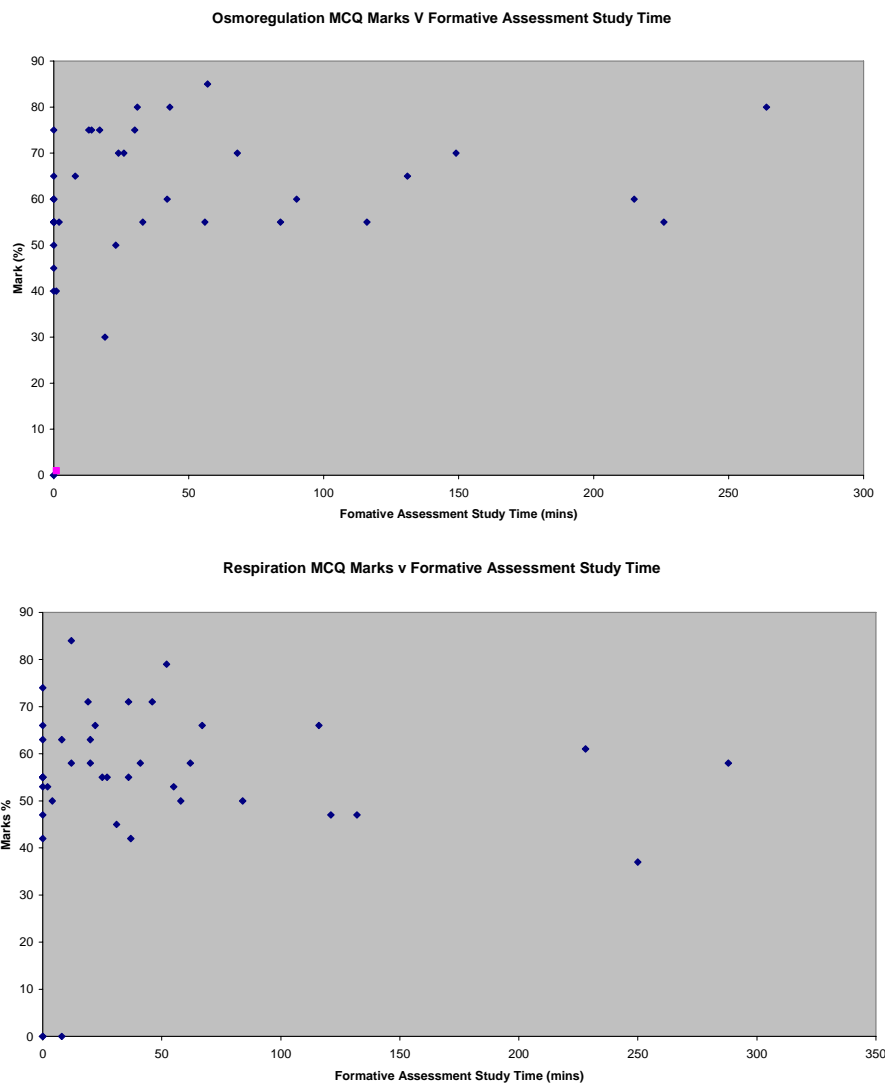
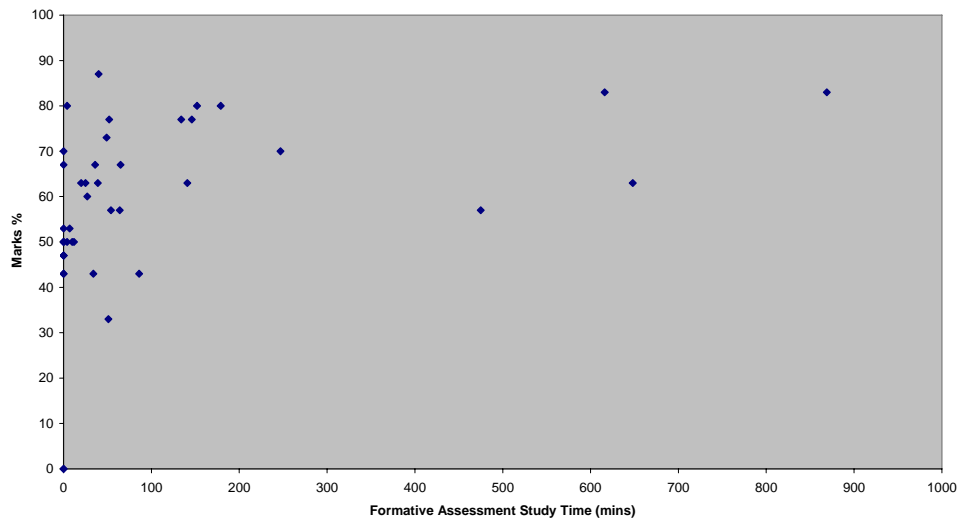
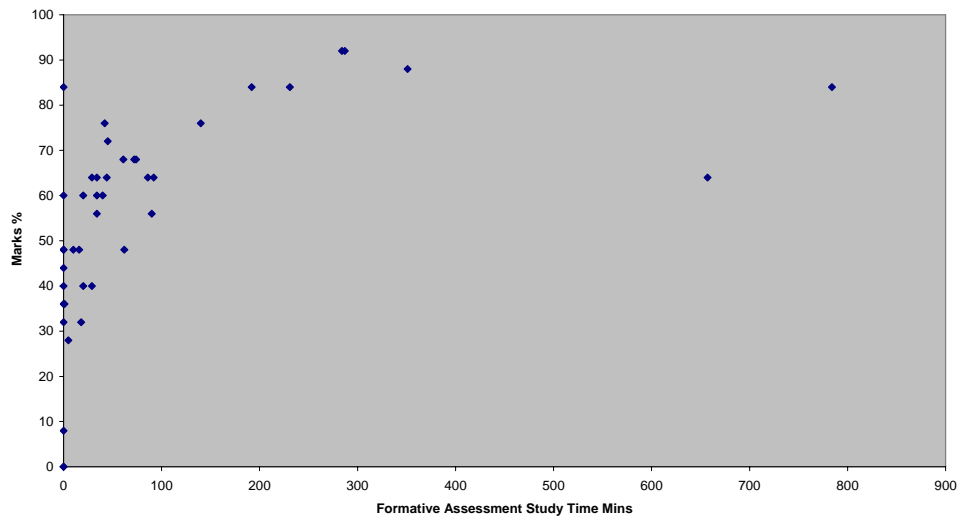


Figure 6

Nerve MCQ Marks v Formative Assessment Study Time 2007=08



Muscle MCQ Marks v Formative Assessment Study Time



Hormone MCQ Marks v Formative Assessment Study Time

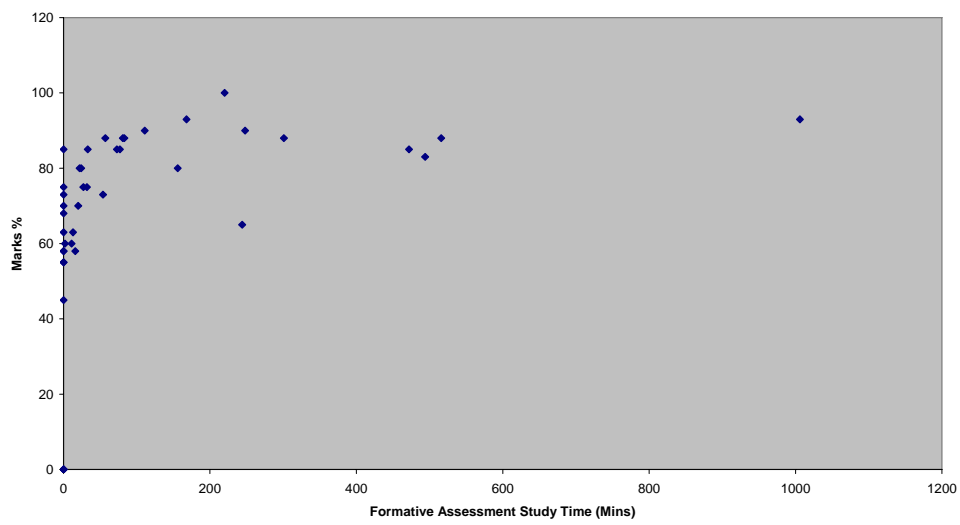


Figure 7

The three graphs in figure 7 above represent the data for the group 2 topics where the formative assessment is by complex quizzes, and all three show clear and statistically significant positive correlations.

Topic	Pearson's R	Probability	N (students)
Nerve Physiology	0.379	p<0.05	42
Muscle Physiology	0.506	p<0.001	42
Hormone Physiology	0.399	p<0.01	42

Table 4

It is concluded that the extra time students spend on the on-line quizzes is more effective than the time spent on the simpler self-tests. This may be because of the multiple attempts at on-line quizzes, each with a different selection of questions, or it may be related to the detailed feedback on correct and incorrect answers. If the formative test gives less feedback and reveals the correct answer at once, there is no evidence that presenting it on-line instead of on paper confers a learning benefit. It should be remembered that although the summative scores for the muscle and hormones topics (table 2) were clearly improved by this regime, but those for the nerve physiology topic were not. This may be related to the fact that the correlation between time spent on this topic and marks obtained were weaker than for the other two (table 4). Nerve physiology is often perceived by students to be conceptually difficult and this topic may need particular care in the writing of questions.

The reason why the examination essay marks were not improved by the same learning regime requires some thought. It may be that the in-course objective questions (both formative and summative) were not closely enough aligned with the end of module examination essays, so providing poor preparation for them. Certainly, no particular thought was given to this either when the objective in-course questions or the examination questions were being written. A related peripheral conclusion from the study is that insufficient attention has been paid to the design of final examination questions, particularly in terms of their depth. This seems to have a major effect on the mean marks, and also on the variance of marks, with the deeper questions resulting in lower mean marks and greater variance. This may mean that they are much better discriminators of student achievement, revealing that a few students develop real insight into the physiological systems being studied but that the majority do not.

There is a great deal more analysis that can be performed particularly on the VLE tracking data, for example comparing the study patterns of successful and unsuccessful students. Do the students make use of the on-line tests in the examination revision period, and if so to what extent? How much use do they make of the on-line answers and feedback for the summative in-course tests? This is a rich mine of data, and it is intended to analyse it further in due course.

Summary

Changing from paper to on-line formative assessment was not shown of itself to improve student performance. But when full advantage was taken of the quiz features in WebCT, student marks on the subsequent summative in-course tests improved, and this was correlated with increased time spent by students on the formative tests. The key features for success were random selection of questions from a sizeable question bank, giving an incentive for repeated attempts at the quiz; and detailed feedback which did not immediately reveal the correct answer. A further conclusion was that if improved performance in the end of module examination is to be achieved, the in-course objective questions need to be more carefully designed as a preparation for it.