Research in Learning Technology
Vol. 21, 2013

COACTION PUBLISHING

Combining the formative with the summative: the development of a two-stage online test to encourage engagement and provide personal feedback in large classes

Susanne Voelkel*

School of Life Sciences, University of Liverpool, Liverpool, UK

(Received 17 July 2012; final version received 21 March 2013)

The aim of this action research project was to improve student learning by encouraging more "time on task" and to improve self-assessment and feedback through the introduction of weekly online tests in a Year 2 lecture module in biological sciences. Initially voluntary online tests were offered to students and those who participated achieved higher exam marks than those who did not, but completion rate was low. Making the tests compulsory led to high completion rates, but class performance decreased, indicating that using the same assessment for formative and for summative purposes is not always beneficial for learning. Finally, these problems were resolved by introducing a two-stage approach: the first stage of each test was formative and provided prompt feedback. However, students had to achieve 80% to progress to the second summative stage of the test. The two-stage online tests led to significantly improved class performance. This novel test design ensures that students go through at least two attempts and therefore fully benefit from the learning opportunities presented by the formative stage. Two-stage online tests present the opportunity to provide regular feedback in large classes and to improve performance not only of good but also of "weak" students.

Keywords: e-learning; e-assessment; action research; higher education

Introduction

Formative assessment and feedback have a powerful influence on student learning (Hattie and Timperley 2007). To be effective, however, feedback needs to be timely and provide information to the student on how to close the gap between current and desired performance (Nicol and Macfarlane-Dick 2006). Growing class sizes make it difficult to offer frequent formative assessments in combination with high-quality feedback. This study reports on the development of a successful, novel, two-stage online test that encourages student engagement and ensures regular, timely feedback to large classes.

Background

We no longer see learning as knowledge acquisition based on teacher transmission, but as a process in which students play an active role in constructing their knowledge

*Corresponding author. Email: svoelkel@liv.ac.uk

Citation: Research in Learning Technology 2013, **21**: 19153 - http://dx.doi.org/10.3402/rlt.v21i0.19153

and skills by building up on prior knowledge and understanding (Palincsar 1998). This concept of "student centred learning" is characterised by active rather than passive learning, deep learning and understanding (rather than surface learning) and an increased responsibility by the student (Biggs and Tang 2007; Lea, Stephenson, and Troy 2003). The process of learning is complex and influenced by students' attributes and experiences as well as by conditions of learning presented by the course (Busato 2000; Harris 1940). There are two factors, however, that are widely seen as hugely important to learning: student engagement and good quality feedback (Gibbs 2010; Trowler and Trowler 2010).

Kuh (2003) describes engagement as "the time and energy students devote to educationally sound activities inside and outside of the classroom" and links it to the "time on task" principle (Chickering and Gamson 1987), holding that "the more students study a subject, the more they learn about it". However, most out-of class learning is allocated to assessed tasks (Innis and Shaw 1997), and Gibbs (2010) argues that assessment has a "profound influence on what, how and how long students study".

Some forms of assessment can instigate inappropriate learning activities. For example, certain forms of multiple-choice tests can lead students to adopt surface rather than deep learning approaches (Scouler and Prosser 1994). Gibbs (2010) suggests that students' approach to learning is more likely to be determined by what students perceive to be the demands of the test, rather than what the teacher actually intended. Good practice, therefore, needs to communicate high expectations (Chickering and Gamson 1987) and assignments need to be perceived as challenging, but possible, to the students. This requires the communication of clear standards and goals. If students don't understand what is expected of them, they tend to revert to surface approach and memorisation (Gibbs 2010). Unfortunately, criteria are seldom meaningful to students and it is often difficult for them to tell what standard is expected (Gibbs and Dunbar-Goddet 2007). The latter authors found that giving out clear goals and standards had little effect on learning, and that it was much more helpful when students received plenty of feedback.

Formative assessment is a form of "assessment that is specifically intended to provide feedback on performance to improve and accelerate learning" (Sadler 1998). Black and William (1998) found that formative assessment can make a strong contribution to the improvement of learning. However, the quality of feedback is of crucial importance. Nicol and Macfarlane-Dick (2006) formulated seven "principles of good feedback practice" proposing that feedback should be prompt, it should clarify what is expected of the student, and it should enable students to improve their performance. Frequent formative assessment with regular high-quality feedback encourages consistent work. Large class sizes, however, make it difficult for teachers to give timely and good quality feedback to frequent assignments. Blended e-learning, which describes a combination of traditional learning with web-based online approaches, offers scope to create additional opportunities for feedback, to ensure immediate feedback, and to help engage students out of class (Sharpe *et al.* 2006). Turney *et al.* (2009) found that blended e-learning can significantly improve student learning provided it is fully aligned to the teaching aims and embedded in the course and that there appears to be a strong positive relationship between the use of e-learning and measures of student engagement (Nelson Laird and Kuh 2005).

Online tests (or e-assessments) as a form of e-learning have become increasingly popular as they are easily accessible to students and can be marked automatically

Citation: Research in Learning Technology 2013; **21:** 19153 - http://dx.doi.org/10.3402/rlt.v21i0.19153

regardless of class size, so that results can be made available immediately (Hepplestone *et al*. 2011). e-assessment is widely accepted by students as part of their university studies and they generally feel that it has a positive impact on their learning (Dermo 2009). The latter author also found, however, that e-assessments using randomly selected questions from a question bank can be perceived by students as unfair. Indeed, Jordan, Jordan, and Jordan (2012) found that different variants of computer-marked questions can behave differently and that it is necessary to monitor performance of supposedly equivalent questions.

Several authors have used e-assessment to good effect in a variety of settings, including weekly online multiple-choice quizzes (e.g. Peat and Franklin 2002), multiple-choice questions with confidence-based marking (Rosewell 2011), and computer-assisted marking of short free-text student responses (Butcher and Jordan 2010). e-assessment usage ranges from purely formative (e.g. Henly 2003) to mainly summative online tests (e.g. Marriott 2009), while Angus and Watson (2009) describe a model that combines formative with summative principles by allowing multiple attempts and using the best attempt for marking purposes. Automatically generated feedback may give information about whether an answer is correct or not, sometimes followed by in-class clarification of common misconceptions (e.g. Hodgson and Pang 2012), or it may be tailored by providing more feedback after each attempt where multiple attempts for each question are allowed (Jordan 2011).

Aim

The aim of this study was to develop and then evaluate the feasibility and effectiveness of weekly online tests in a Year 2 theory module in biological sciences. The tests were intended to encourage student engagement with the lecture material, and to support their learning through formative assessment and feedback. The study was conducted as an action research project. Three cycles were completed, in which an online test design was introduced, evaluated and reflected upon and then adapted accordingly in the next cycle. In the first cycle, voluntary online tests were introduced, which were then made compulsory in the second cycle. Cycle 3 saw the introduction of a two-stage online test approach with an initial, purely formative stage, followed by a second, summative stage that could only be accessed after 80% were reached in the first stage.

Research design

The study was designed in the form of an action research project in which three cycles were completed. Action research combines various stages which include the identification of a problem or question (How can I improve engagement and learning?), the process of tackling the problem (interventions: introduction of online tests), evaluation and reflection, followed by further modification of practice (e.g. Norton *et al*. 2001). The general hypothesis for the project was that students will enhance their learning following the introduction of online tests.

The subject of this study was an animal physiology module, which is offered to second year students from various programmes within biological sciences at the University of Liverpool. This theory module is taught through 18 lectures in six weeks (three lectures per week). Before the beginning of this study, the course was assessed solely by a final exam, which took place about three months after the last

lecture. The final exam consisted of three parts: (1) short answer questions, (2) disclosed essay, and (3) unseen essay. Formative assessment was only provided in the form of in-class activities where volunteers answered one or two questions per lecture. The three cycles of the project took place in the academic years 2008/2009 (Cycle 1), 2009/2010 (Cycle 2) and 2010/2011 (Cycle 3). Class size and gender composition of classes in each cycle are shown in Table 1.

Evaluation of the outcome

Student performance

Marks from all assignments were collected and analysed. Average marks from previous cohorts were included to compare overall performance before and after the implementation of the interventions. In addition, mean marks of other second year theory modules were used to compare performance of the same cohort of students within a variety of learning environments. To assess the magnitude of any significant changes following the interventions, effect sizes were calculated according to Fan (2001) by using Cohen's *d* (Equation 1), which is based on standardized group-mean differences.

$$\text{Equation 1: } d = \frac{X1 - X2}{\text{SD}_{\text{pooled}}}$$

(*X1* and *X2* are the average performance marks for Group 1 and Group 2 (for example, average class mark from Cycle 1 as compared to Cycle 2), and $\text{SD}_{\text{pooled}}$ is the pooled standard deviation between the two groups.)

Statistical analysis of performance data was done using Sigma Plot (version 11.1). Statistical differences between exam average marks were assessed using one-way ANOVA or Student's *t*-test for independent samples, as appropriate. Significance was assessed at the $p < 0.05$ level.

Student evaluations

Students' views were gauged after Cycle 2 using a paper-based questionnaire and a focus group with 10 volunteers. Cycle 3 was evaluated through three electronic surveys during and after the course. Questionnaires contained a mixture of mainly Likert-style fixed-answer questions and free-text comments. In addition, students

Table 1. Class size, percentage of male and female students, mean module mark and standard deviation, percentage of failed students, and online test completion rate.

Year	Class size	Male (%)	Female (%)	Mean mark (%)	SD	Significance	No. fails (%)	Online test completion rate (%)
2006/2007	137	38	62	57	14	a	10	–
2007/2008	108	38	62	53	15	a, c	18	–
Cycle 1 2008/2009	83	43	57	60	13		6	33
Cycle 2 2009/2010	91	45	55	57	11	a	5	98
Cycle 3 2010/2011	78	33	67	64	14		8	99

For comparison, the 2 years before the study are included. Significant differences: a = different to 2010/11, c = different to 2008/9.

were invited to comment on the module in a school-wide end-of-module evaluation. Free-text comments from questionnaires as well as the focus group transcript were analysed using thematic analysis (Braun and Clarke 2006).

Before taking part in the evaluations, students were informed verbally and in writing that: (1) participation was completely voluntary and that they could stop participating at any time, (2) the surveys were anonymous and that participants could not be identified by the answers, (3) the results might get published, and (4) they could contact the author with questions at any time. To ensure anonymity for focus group participants, a member of staff from a different department facilitated the focus group, and people from outside the department undertook transcripts of the recordings. Students were informed that the focus group discussions were recorded, and they all gave their permission beforehand.

The project

The following section provides for each of the three cycles a description of the intervention, an analysis of the outcome and the results of student evaluations, and finally a reflection on the success of the intervention, which then leads to a modification of the following cycle. This is followed by a discussion of additional factors that may have influenced the outcome of the study.

Cycle 1 (2008/9)

The intervention

In the first cycle of the project, online tests were introduced on a voluntary basis, to offer students the opportunity for self-assessment. A total number of six tests, each consisting of a set of six to eight fixed-answer-type questions, were posted on the university's virtual learning environment (Blackboard) and were available throughout the course. There was a variety of question types from multiple choice (single best answer out of a number of options), multiple answers (several correct choices out of a number of options), ordering (where students have to select the correct order of a series of items), matching (requires pairing of items), calculated formula (these questions contain a formula and the variables can be set to change for each student), calculated numeric (students have to enter a number as an answer) and hotspots (students indicate an answer by clicking on a specific area in a diagram). All of these were fixed-answer type questions that allowed automatic marking. The variety of question types was chosen to promote understanding and application rather than memorization. The test topics were closely related to the lecture topics. Students had multiple attempts and no marks were attached. After submission of each test, students immediately received their scores and information on the correctness of their answers.

Analysis

After implementation of the voluntary online tests, average exam performance increased significantly from 53 to 60% (Table 1), which corresponds to an effect size of 0.5. These data suggest that the introduction of voluntary online tests helped improve student performance. It has to be noted, however, that the average exam

performance in this year was not significantly different from two years ago, in 2006/7 (Table 1). Nevertheless, there seems to be a beneficial effect attached to the completion of the online tests. Figure 1 shows that students who did complete the online tests were about three times as likely to get a first class mark and less than half as likely to achieve a II.2 or below than those who did not attempt the online tests. The average exam results of those who attempted the tests was significantly higher ($65 \pm 11\%$) than of those who did not ($58 \pm 13\%$) (t-test, $p = 0.017$). Kibble (2007) also found a significantly better exam performance in students who participated in formative, voluntary online quizzes, compared to those who did not. These results could indicate that completing the formative online tests improved student learning. However, they could also be due to the fact that students with a more successful approach to studying were more likely to complete voluntary work. Henly (2003) showed that high-performing students were twice as likely to access formative online assessment than low-performing students. In the present study, only about a third of the class completed the tests, a result that is comparable to other studies (e.g. Kibble 2007), so even if there was a benefit to be gained, only a minority of the class had benefitted from it.

On reflection

Providing the voluntary online tests was technically relatively easy, and did hardly require any teacher input once the tests were made available. However, a lot of time went into the design of the questions to make sure that these were relevant and testing for higher cognitive skills such as understanding and application. Various authors have demonstrated that although it takes more time and effort, fixed-answer questions can be constructed in a way that they test understanding and application in addition to knowledge (e.g. Butcher 2008; Hampton 1993). The increase in the overall exam results was encouraging, but not conclusive to show that the tests were beneficial, and the fact that only a third of the class attempted them was worrying.

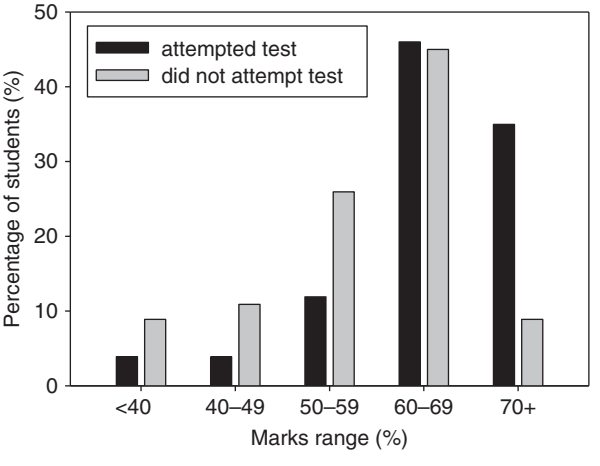

Figure 1. Percentage of students who have or have not attempted the voluntary online tests, within their achieved exam mark ranges.

Cycle 2 (2009/10)

The intervention

Following the analysis and evaluation of the first cycle, the online tests were made compulsory to improve completion rate. The six tests now contributed 20% towards the final module mark (i.e. each test contributed 3.33%). The final exam remained unchanged but now counted as only 80% towards the final mark. Tests were released weekly during the lecture period of six weeks and the students had one week to complete each set. The question sets again contained a variety of question types as in Cycle 1, but now consisted of 8–12 questions each. The question settings allowed students to do each test only once. All question sets were closely linked to the topics that were discussed in the lectures within the same week. All students received the same questions, with the exception of calculated formula questions that were based on randomly allocated variables for each student. After submission of the test students got their overall score, but initially no information about which answers were right or wrong was revealed. The initial withholding of information on correctness was deliberately chosen to discourage collusion. However, general feedback was given in the lecture following the submission date, and correct/incorrect answers were released one week after the test submission deadline when no more marks could be gained by late submission.

Analysis

Making the online tests compulsory succeeded in ensuring a high participation rate (98%). However, the average module mark went down from 60 to 57%, and was no longer statistically different from the two years before the study (Table 1). This trend was mirrored by the total average exam mark, and also Part A and Part B of the exam (Figure 2).

Student evaluation

A paper-based questionnaire was handed out in the last lecture and completed by 31 students (34% of the class). The answers indicated that students had an overall positive outlook regarding the online tests (Figure 3). Around 80% thought the instructions were clear, the tasks relevant, and that lectures prepared them well for the tests. More than 60% enjoyed doing the tests, while less than 30% found them tedious and boring. Over 90% found the tests helped their learning. However, less than 60% found the feedback for the tests (which was given in the lecture in the form of general class feedback as well as individual feedback one week after each submission deadline) helpful for their next test. On average, 65% of the students spent at least one hour per week on the online tests (data not shown). It has to be noted that this questionnaire was only submitted by one third of the class, and the results may not be representative for the whole class.

Students' views were also gauged from a focus group consisting of 10 students and from online module evaluations where 39 students commented. Thematic analysis of all comments addressing online tests confirmed that generally students liked the online tests (Table 2). They appreciated that their marks were not wholly based on the final exam, and they also liked the variety of the assessment questions. Some said they liked the fact that they could do the test whenever it suited them, and

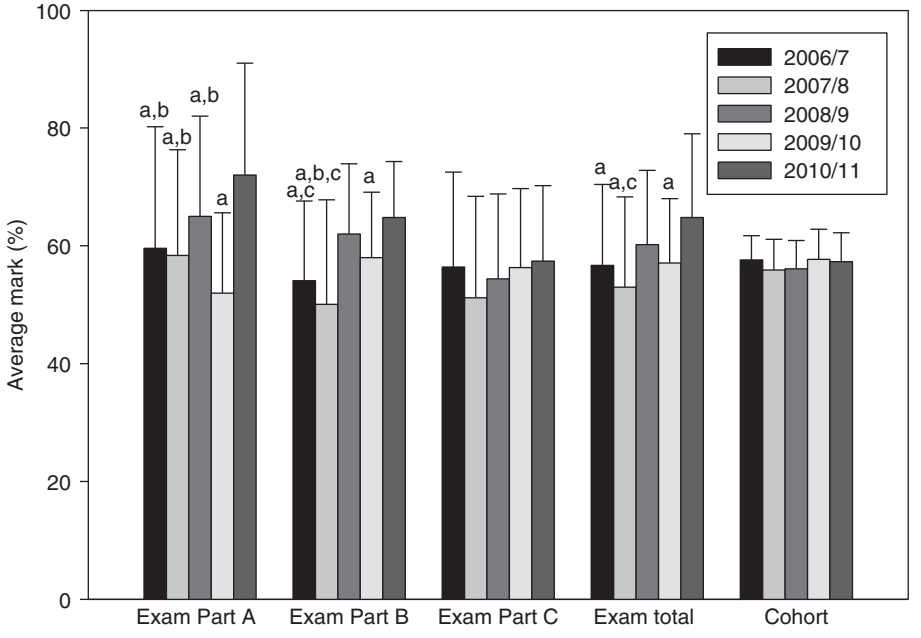

Figure 2. Average exam marks ± standard deviations (%) for exam Parts A, B and C and the total exam. For comparison, the average Year 2 theory marks for the same cohorts are shown (a, b, and c = significantly different from 2010/11, 2009/10, and 2008/9, respectively. Part B of the exam in 2010/11 was replaced by a homework essay).

they found the tests to be useful as "refreshers". Some students said that the tests made them look at their lecture notes again, something they would not have done otherwise until shortly before the exam. Some also appreciated the feedback they were getting on their understanding. Some problems, however, emerged: students did not like the fact that they were not getting immediate information about which answers were right or wrong, they would have liked more personalized feedback.

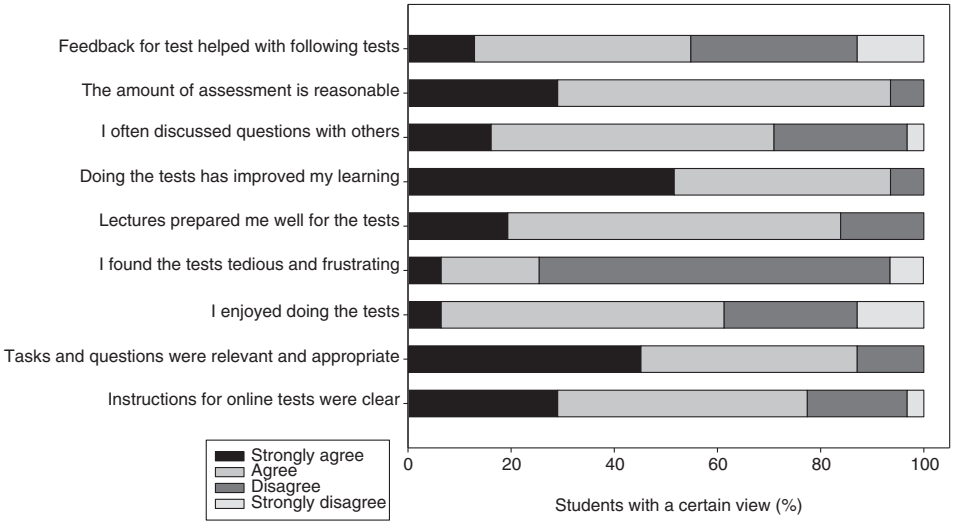

Figure 3. Results of the end of Cycle 2 questionnaire ($N = 31$).

Table 2. Themes and sub-themes identified from the analysis of the students' views on the online tests in Cycle 2.

Theme	Sub-theme	Typical comments
Learning	Self-assessment	Focus group: "it shows you what you probably need to revise more for the exam like, I did quite bad on one of them, so I know that I should work on that for the exam" Evaluation (3): "The online assessments were very good as they allowed you to see how well you understood the material as you went along".
	Engagement	Focus group: "It makes you concentrate on your work, like actually go over your notes", "half the time people just go to lectures and don't do anything else until the end. A lot of people actually do that, so if you do have the assessment, yeah, it sounds like a lot more work, but at least it's making you do the work" Evaluation (5): "I feel the online work was good because it forced students to engage themselves in the material", "tests helped to strengthen learning".
	Feedback	Focus group: "I think the problem was, you got your mark back and you didn't know what you got wrong", "Yeah, because then (if the system told you what was wrong) you'd actually learn from what you got wrong, whereas now we've done the assessment and then it's like, okay, don't think about it" Evaluation (8): "the self-assessment questions which are worth 20% of this module were very confusing and it would have been better if they would give us the answers as we went along, as we could not see where we'd made our errors".
Procedure	Questions	Focus group: "I just didn't know what on earth it was asking, because it was really badly worded", "Yeah, well, it's like I know the stuff, so I'll put down the right answer, but the computer's marking it wrong". Evaluation (8): "The MCQ assessment questions could have been a little clearer", "the computer did not recognize the right answers and gave different marks to people who I knew had put more or less the same answers down as it depended on where they clicked on the diagram".
	Timing	Focus group: "I think people like the flexibility", "it's probably more sense to do by 4 pm on a Monday" Evaluation (3): "Longer deadlines for self-assessment tests, was very had to complete in 3 days if you were away/ working at the weekend".
	Weighting	Evaluation (9): "I like that this module is not 100% exam based as it reduced the pressure around exam time".

No attempt was made to quantify the results of the focus group as the recording did not allow to distinguish between different speakers and it was not possible to tell whether an issue was brought up several times by the same speaker or by several different speakers. In the module evaluation, however, it was possible to distinguish between different individuals. Number of comments referring to a certain sub-them is given in brackets.

Some students had problems understanding what they were asked to do, and felt that the computer marked their answer wrong. Finally, a few said that they needed more time to complete the tests.

On reflection

It is not surprising that making the tests worth a considerable amount of credits achieved a high completion rate. Kibble (2007) also found that incentives in the form of credits increased participation. However, despite the fact that now almost all students completed the tests, overall exam performance dropped to the same

standard as before the start of the project. An explanation for this failure of the compulsory online tests to improve student performance may be found within the evaluation results. On the one hand, students appreciated that the tests gave them opportunity for self-assessment and also forced them to go over their notes. In that sense, the compulsory online tests did what was intended: to increase students' "time on task". However, students criticized the lack of prompt and personal feedback and they felt that learning would have been more successful if they had received immediate information on what exactly they got wrong. Some students even found the tests confusing when they did not understand why they got low marks. These comments confirm that the general feedback in the lectures was not sufficient in giving students the information they needed to improve their learning and understanding, and most students did apparently not take the opportunity to access their individual feedback that was released after a week. Hattie and Timperley (2007) emphasize that effective feedback needs to provide information that specifically relates to the task, so that students can develop self-regulation and error detection strategies and use the feedback to then tackle more challenging tasks. Scores alone do not provide the necessary information for this (Gipps 2005). Furthermore, knowing which answers were correct is just as important as knowing which answers were wrong (Hattie and Timperley 2007). The setup in this cycle of the project initially provided scores only and general feedback in class obviously did not provide enough information to the students to be able to close the gap between current performance and the goal (Nicol and Macfarlane-Dick 2006). The summative assessment might have superseded the formative role, a setup that has been found to be less effective for learning than formative assessment alone (Miller 2009).

Cycle 3 (2010/11)

The intervention

After evaluation and reflection on the outcome of the second cycle, the weekly online tests were modified to allow more immediate and personal feedback. Each test was divided into two stages: a purely formative stage A, and a summative stage B (Figure 4). As in Cycle 2, both test stages comprised of various types of fixed-answer questions and addressed the topics that were discussed in the lectures during each respective week. The questions in stage B were variants of the questions in stage A. The question settings initially allowed students only to see and access the formative stage A. Students could repeat stage A as often as they liked and upon completion of each attempt they would see which answers they got right and which were wrong. Questions that were answered wrong triggered specific feedback containing hints and tips to help answer the question. The marks for stage A did not count for the final module mark. However, students had to achieve at least 80% in stage A to be able to see and access the summative stage B of the test. The marks from stage B did count towards the final mark (see Table 3 for details on the test regime). Students could do stage B only once, and initially they only got their score back, but right and wrong answers were revealed one week after the submission deadline. A month before the exam the test settings were altered so that students could voluntarily redo all tests for exam revision. Other interventions included a homework essay replacing the disclosed final exam essay, and a mobile phone-based audience response system described elsewhere (Voelkel and Bennett 2013) which was used for in-class activities

Citation: Research in Learning Technology 2013; **21:** 19153 - http://dx.doi.org/10.3402/rlt.v21i0.19153

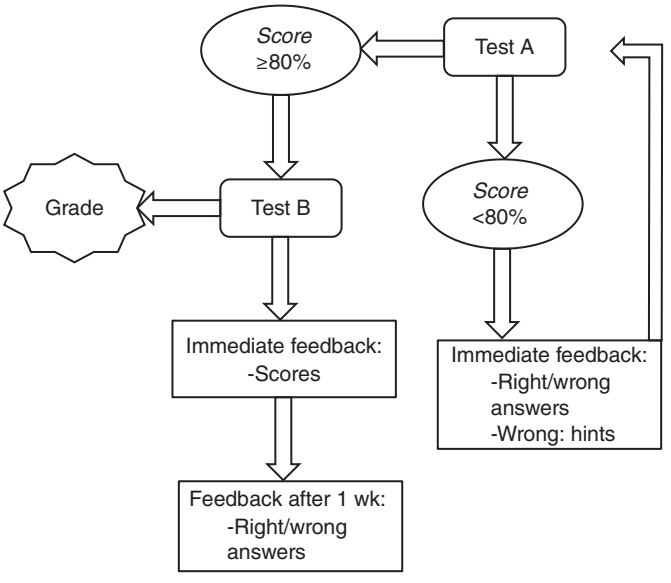

Figure 4. The two-stage online test design. Start with Test A.

that were previously undertaken by relying on individual volunteers answering the questions.

Analysis

Similar to Cycle 2, completion rate was very high (99%) as the online tests again were worth 20% of the module mark. The average module mark went up to 64%, which

Table 3. Test schedule for the first 3 weeks of the module.

	Week 1		Week 2		Week 3	
Monday	Lecture 1	Test 1 (A+B) available	Lecture 4	Test 1 B submission deadline (Scores) Test 2 (A+B) available	Lecture 7	Test 1 B (Full feedback) Test 2 B submission deadline (Scores) Test 3 (A+B) available
Tuesday						
Wednesday						
Thursday	Lecture 2		Lecture 5		Lecture 8	
Friday	Lecture 3		Lecture 6		Lecture 9	

For example, Test 1 covered topics from Week 1 (Lecture 1, 2 and 3) and Part A was available from Monday Week 1. Test 1 A could be done multiple times, and detailed feedback for incorrect answers was given immediately after each attempt. Once students achieved at least 80% in Test 1 A, Test 1 B became available. The deadline for submission for Test 1 B was Monday Week 2. Initially, only scores were given. One week after the deadline (i.e. Monday Week 3), full feedback was available for Test 1 B in the form of correctness of individual questions, and revealing correct answers.

was significantly higher than the marks in Cycle 2 (effect size 0.6), and both years before the study, but not significantly different from Cycle 1 (Table 1). These differences are reflected in the exam performance (not including the online test marks) as well as in Part A of the exam (Figure 2). Part B performance cannot directly be compared with previous years because the disclosed essay was replaced by a homework essay. The unseen exam essay (Part C) mark, however, did not differ from previous years (Figure 2).

Student evaluation

Students' views were evaluated four times at different points in the course. An online survey took place in Week 3 ($n = 75$, 95% of the class). A second survey was performed in Week 5 ($n = 64$, 82%), and a third survey directly after the exam ($n = 47$, 60%). Finally, an end-of the module questionnaire offered the option of commenting on the whole course.

The first survey showed that more than 80% of the class found that the online tests helped their understanding, and they also found the feedback they got for the first part of the test helpful (Figure 5). Around 75% agreed that the questions were clear, and that the first test helped them get better marks in the second test. Just over 60% found the tests moderately difficult and 25% fairly difficult and for the first two tests 55 and 70% of the students, respectively, needed more than one attempt to gain 80% of the marks (data not shown).

In the second survey, more than 80% agreed that the weekly online tests gave them plenty of feedback on their learning and that the tests made them look at lecture material more than they would otherwise have done (Figure 6). About 70% were planning to redo the tests for exam revision. Around 60% felt more confident about the exam because of the tests, and said that they usually look at the feedback for the second test which was revealed a week after the deadline. The third survey concentrated on how useful the students found the online tests for their exam revision. 40% of the respondents actually repeated the online tests as part of their revision, and another 40% included the feedback they got. Over 90% agreed that they found the weekly online tests very useful (60%) or somewhat useful (36%) for their learning (data not shown).

To evaluate the free-text comments from all three surveys and the end-of-module questionnaire, all comments that addressed online tests were analysed. Two main themes were identified: learning and experience, which were then further divided into sub-themes (Table 4). Many comments addressed the fact that the tests increased their "time on task", making them "go over their notes", and many also found that the tests helped them to understand the material better. Students also appreciated the opportunities for self-assessment and the frequent feedback. A lot of students valued that the continuous assessment "took pressure off the exam".

On reflection

In the third cycle of the project, much more emphasis was put on the formative aspect of the online tests. This was done by using a novel two-stage test approach where formative and summative parts of the tests were completed separately, but were linked through the 80% rule. This approach seemed to have worked very well.

Citation: Research in Learning Technology 2013; **21:** 19153 - http://dx.doi.org/10.3402/rlt.v21i0.19153

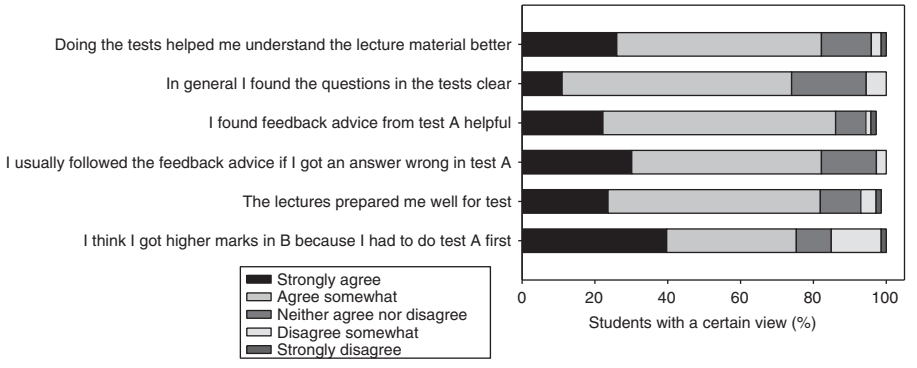

Figure 5. Results of a questionnaire completed halfway through Cycle 3 ($N = 74$).

The summative part of the online test ensured a high completion rate, whereas the first, formative part gave students plenty of opportunity for self-assessment and provided prompt, detailed feedback that gave students information on what they needed to do to improve their performance. These results suggest that separation of formative and summative online tests was successful and avoided the problems reported by Miller (2009). The observed overall effect size of 0.6 is comparable to a study by Hattie and Timperley (2007) summarizing effect sizes relating to various feedback methods. According to their study "computer-assisted instructional feedback" had an effect size of 0.5. The significant increase in student performance of this size following the introduction of a two-stage online test is, therefore, meaningful, both in a statistical as well as in a practical way (Fan 2001).

It is also obvious that students rated the two-stage online tests very highly and appreciated the effect they had on their learning. Results from the student questionnaires signify a much higher satisfaction with feedback for the test in comparison to Cycle 2. Presumably, the two-stage tests increased the "time on task" even further than did the one-stage tests in Cycle 2. More importantly, there is good evidence that students spent this additional time on educationally meaningful tasks. Not only do they tell us that they spent more time revisiting their lecture notes, but the 80% rule also encouraged them to act on the immediate feedback they received if they got questions wrong (Figure 5). Similarly, Butcher, Swithenby, and Jordan (2009) suggest that interactive computer-assisted assessments allowing multiple

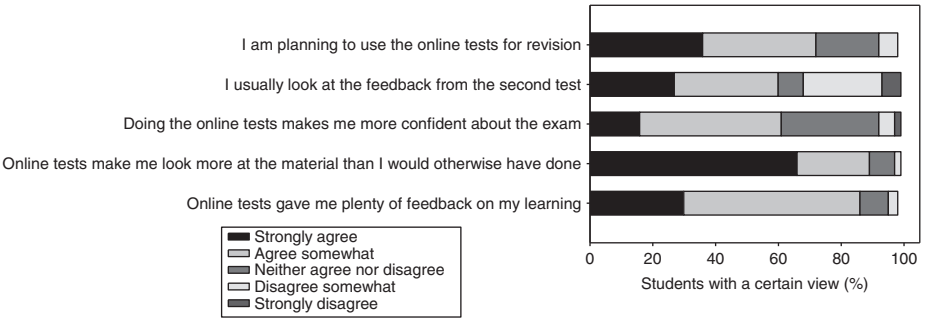

Figure 6. Results of a questionnaire handed out at the end of Cycle 3 ($N = 64$).

attempts with built-in feedback can engage students in meaningful learning activities, asking them to act on feedback "there and then".

Other factors that might have an impact on the results

Cohort differences

To see whether any differences in the performance of students in the study module were due to variations in cohort aptitude, average marks of other theory modules taken by the study cohort were analysed. There was no significant change in the average cohort performance between 2007 and 2011, indicating that any changes in the study module performance were not due to an unusual cohort.

Gender and class size

Class size in this module changed from over 100 in the two years before the start of the project, to about 80–90 during the three years of the study (Table 1). However, no significant relationship between class size and performance was found.

The percentage of female students in the module was always higher than that of the male students, and the difference was particularly high in the third cycle when more than twice as many female students took the class than male students (Table 1). There has been some discussion about gender-based performance with some studies indicating an overall higher achievement in women as compared to men (McNabb, Pal, and Sloane 2002). The differences in performance across the years may therefore

Table 4. Views on online tests in Cycle 3. Numbers in brackets give the number of times an item was mentioned.

Theme	Sub-theme	Typical comments
Learning	Time on task (14)	"The continuous assessment throughout the module meant that I was forced to go over lectures from that week and make sure I understood them", "The online tests really made me go over my notes"
	Self-assessment and feedback (10)	"I really appreciate the tests. My first attempt is usually quite bad and after repeating the tests several times to get above 80% I find I slowly start to learn the content", "The assessments throughout the course were very helpful to see how much I understood the material"
	Helps understanding and revision (13)	"They helped me understand the lecture information in more detail", "The VITAL tests are a really good way of revising"
Experience	Enjoyment (3)	"I really like these tests", "I actually enjoy them"
	Takes pressure off exam (14)	"It takes a bit of pressure off the final exam"
	Problems (8)	"Maybe introduce some long answer questions to help prepare for the exam essay. I find it very difficult to write essays about information I have been taught/ learnt as bullet points.", "It is sometimes confusing with the tick box questions when there is no definite amount to tick, it would be more helpful if the questions stated tick (numbers) of boxes. I understand however for some questions this would be detrimental to the learning".

Citation: Research in Learning Technology 2013; **21:** 19153 - http://dx.doi.org/10.3402/rlt.v21i0.19153

have been affected by gender composition. No significant differences between female and male students' exam performance was seen (Figure 7). Male performance did not show any significant changes during the years. Female students' exam marks, however, increased significantly between Cycle 2 and Cycle 3 (effect size 0.9) and between 2008 and the first cycle of the study (effect size 0.6). These data may indicate that female students benefit more from the online tests than male students.

Other interventions

Other changes included the introduction of phone polls (an audience response system similar to "clickers", but using students' mobile phone devices) for in-class self-assessment in Cycle 3. This innovation was very well received by the students as they enjoyed the fact that they could participate anonymously and it introduced breaks in the lecture (see Voelkel and Bennett 2013). It is unlikely, however, that the phone polls had a great impact on class performance in this module because similar class activities (without polling) had been used already in Cycle 1 and Cycle 2, the only difference being that before the phone polls answers were volunteered by individual students. Another change was the introduction of a homework essay assignment, which replaced the disclosed essay in the exam. The marks for the homework essay in Cycle 3 are significantly higher than the marks for the disclosed exam essay in most previous years (Figure 2). This is not surprising, as continuous assessment tasks usually result in better marks than exam work (Bridges *et al*. 2002). One could have expected that the additional essay writing practice, in combination with feedback would lead to an improved performance in the unseen exam essay (Part C), but unfortunately this was not the case (Figure 2).

However, marks from Part A of the exam (the short answers), were significantly higher than in previous years. In fact, the effect size of Part A alone was 1.2, the highest increase in marks seen in any of the components. The performance in exam Part A is less affected by writing competence and requires students to analyse diagrams, identify structures, choose the correct formula for numerical questions and apply numerical skills. All of these skills were previously practiced in the online tests

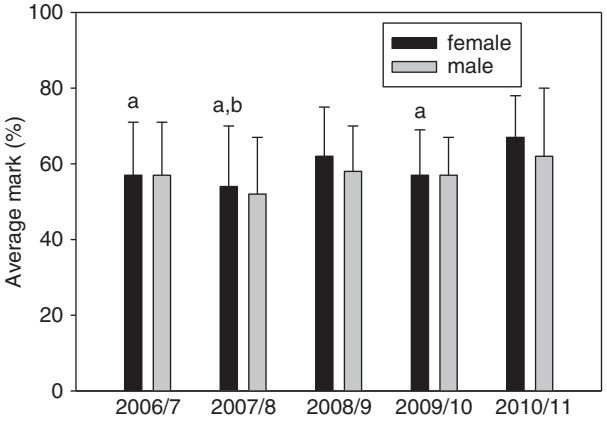

Figure 7. Average exam marks ± standard deviation (%) of male and female students (a, b = significantly different from 2010/11 and 2009/10, respectively).

and it is, therefore, highly likely that the weekly online tests had more of an impact on Part A performance than the additional essay writing assignment.

Conclusion

This project aimed to improve student learning by introducing online tests, which were meant to engage students, increase the time they spend out of class on educationally meaningful activities, and to provide opportunities for self-assessment and feedback. The results suggest that increasing the time on task alone (by forcing them to spend time on online tests) did not improve student learning. Only when students were guided towards a meaningful interaction with the material, learning (as measured by exam performance) improved. The prompt, specific feedback after the formative part of the online tests enabled the students to see exactly what they needed to do in order to improve their performance. Students need to make sense of what they have learnt before they are ready to move on. Giving feedback to incorrect answers and confirming correct answers contributed towards empowering students to take responsibility for their own learning (Hattie and Timperley 2007).

Several previous studies have reported on online test designs that foster student engagement by encouraging multiple attempts and providing immediate feedback (e.g. Jordan 2011; Marriott 2009; Peat and Franklin 2002). However, some of these may allow an inappropriate use of the quizzes. Kibble (2007) found that in a test setting where marks were awarded for the best out of two attempts, a significant number of students scored highly on their first attempt and then did not take a second quiz, thereby missing a learning opportunity. These students often could not sustain their high performance in the summative assessment. In the novel two-stage online test design presented in this study students have to take at least two attempts (stage A and stage B), but in reality often three or more, to achieve any marks at all. This test design, therefore, has the potential to significantly improve learning in classes of all sizes and can be a valuable tool for practitioners in a variety of disciplines.

Acknowledgements

I would like to thank Peter Kahn and Tunde Varga-Atkins (Centre for Lifelong Learning, University of Liverpool) for advice during the project and for reading and commenting on the manuscript, respectively.

References

Angus, S. D. & Watson, J. (2009) 'Does regular online testing enhance student learning in the numerical sciences? Robust evidence from a large data set', *British Journal of Educational Technology*, vol. 40, no. 2, pp. 255–272.

Biggs, J. & Tang, C. (2007) *Teaching for Quality Learning at University*, Open University Press, McGraw-Hill Education, Maidenhead, Bershire, UK.

Black, P. & William, D. (1998) 'Assessment and classroom learning', *Assessment in Education: Principles, Policy & Practice*, vol. 5, no. 1, pp. 7–74.

Braun, V. & Clarke, V. (2006) 'Using thematic analysis in psychology', *Qualitative Research in Psychology*, vol. 3, pp. 77–102.

Bridges, P., *et al*. (2002) 'Coursework marks high, examination marks low: discuss', *Assessment & Evaluation in Higher Education*, vol. 72, pp. 35–48.

Busato, V. V., *et al*. (2000) 'Intellectual ability, learning style, personality, achievement motivation and academic success of psychology students in higher education', *Personality and Individual Differences*, vol. 29, no. 6, pp. 1057–1068.

Butcher, P. G. (2008) 'Online assessment at the Open University using open source software: Moodle, openmark and more', *CAA Conference*, Loughborough, UK, pp. 1–12.

Butcher, P. G. & Jordan, S. E. (2010) 'A comparison of human and computer marking of short free-text student responses', *Computers & Education*, vol. 55, no. 2, pp. 489–499.

Butcher, P. G., Swithenby, S. J. & Jordan, S. E. (2009) 'e-Assessment and the independent learner', *ICDE World Conference on Open Learning and Distance Education*, Maastricht, The Netherlands, pp. 1–8.

Chickering, A. W. & Gamson, Z. F. (1987) 'Seven principles for good practice in undergraduate education', *The American Association for Higher Education Bulletin*, vol. 40, pp. 3–7.

Dermo, J. (2009) 'e-Assessment and the student learning experience: a survey of student perceptions of e-Assessment', *British Journal of Educational Technology*, vol. 40, no. 2, pp. 203–214.

Fan, X. (2001) 'Statistical significance and effect size in education research: two sides of a coin', *Journal of Educational Research*, vol. 94, pp. 275–282.

Gibbs, G. (2010) *Using Assessment to Support Student Learning*, L.M. University, Leeds.

Gibbs, G. & Dunbar-Goddet, H. (2007) 'The effects of programme assessment environments on student learning', *The Higher Education Academy*, [online] Available at: http://www.heacademy.ac.uk/assets/documents/teachingandresearch/gibbs_0506.pdf

Gipps, C. V. (2005) 'What is the role for ICT-based assessment in universities?', *Studies in Higher Education*, vol. 30, no. 2, pp. 171–180.

Hampton, D. R. (1993) 'Textbook test file multiple-choice questions can measure (a) knowledge, (b) intellectual ability, (c) neither, (d) both', *Journal of Management Education*, vol. 17, no. 4, pp. 454–471.

Harris, D. (1940) 'Factors affecting college grades: a review of the literature, 1930–1937', *Psychological Bulletin*, vol. 37, no. 3, pp. 125–166.

Hattie, J. & Timperley, H. (2007) 'The power of feedback', *Review of Educational Research*, vol. 77, no. 1, pp. 81–112.

Henly, D. C. (2003) 'Use of web-based formative assessment to support student learning in a metabolism/nutrition unit', *European Journal of Dental Education*, vol. 7, pp. 116–122.

Hepplestone, S., *et al*. (2011) 'Using technology to encourage student engagement with feedback; a literature review', *Research in Learning Technology*, vol. 19, no. 2, pp. 117–127.

Hodgson, P. & Pang, M. Y. C. (2012) 'Effective formative e-assessment of student learning: a study on a statistics course', *Assessment & Evaluation in Higher Education*, vol. 37, no. 2, pp. 215–225.

Innis, K. & Shaw, M. (1997) 'How do students spend their time?', *Quality Assurance in Education*, vol. 5, no. 2, pp. 85–89.

Jordan, S. (2011) 'Student engagement with assessment and feedback: some lessons from short-answer free-text e-assessment questions', *Computers & Education*, vol. 58, no. 2, pp. 818–834.

Jordan, S., Jordan, H. & Jordan, R. (2012) 'Same but different, but is it fair? An analysis of the use of variants of interactive computer-marked questions', *International Journal of eAssessment*, vol. 2, no. 1, [online] Available at: http://oro.open.ac.uk/33705/1/Jordan_Sally_variants_2011.pdf

Kibble, J. (2007) 'Use of unsupervised online quizzes as a formative assessment in a medical physiology course: effects of incentives on student participation and performance', *Advances in Physiology Education*, vol. 31, pp. 253–260.

Kuh, G. D. (2003) 'What we're learning about student engagement from NSSE: benchmarks for effective educational practices', *Change: The Magazine of Higher Learning*, vol. 35, no. 2, pp. 24–32.

Lea, S. J., Stephenson, D. & Troy, J. (2003) 'Higher education students' attitudes to student-centred learning: beyond 'educational bulimia?', *Studies in Higher Education*, vol. 28, no. 3, pp. 321–334.

Marriott, P. (2009) 'Students' evaluation of the use of online summative assessment on an undergraduate financial accounting module', *British Journal of Educational Technology*, vol. 40, no. 2, pp. 237–254.

McNabb, R., Pal, S. & Sloane, P. (2002) 'Gender differences in educational attainment: the case of university students in England and Wales', *Economica*, vol. 69, no. 275, pp. 481–503.
Miller, T. (2009) 'Formative computer-based assessment in higher education: the effectiveness of feedback in supporting student learning', *Assessment & Evaluation in Higher Education*, vol. 34, no. 2, pp. 181–192.
Nelson Laird, T. F. & Kuh, G. D. (2005) 'Student experiences with information technology and their relationship to other aspects of student engagement', *Research in Higher Education*, vol. 46, no. 2, pp. 211–233.
Nicol, D. J. & Macfarlane-Dick, D. (2006) 'Formative assessment and self-regulated learning: a model and seven principles of good feedback practice', *Studies in Higher Education*, vol. 31, no. 2, pp. 199–218.
Norton, L. S., *et al*. (2001) 'The pressures of assessment in undergraduate courses and their effect on student behaviour', *Assessment & Evaluation in Higher Education*, vol. 26, no. 3, pp. 269–284.
Palincsar, A. S. (1998) 'Social constructivist perspectives on teaching and learning', *Annual Review of Psychology*, vol. 49, pp. 345–375.
Peat, M. & Franklin, S. (2002) 'Supporting student learning: the use of computer-based formative assessment modules', *British Journal of Educational Technology*, vol. 33, no. 5, pp. 515–523.
Rosewell, J. P. (2011) 'Opening up multiple-choice: assessing with confidence', *CAA Conference: Research into e-Assessment*, Southampton, UK, pp. 1–3.
Sadler, D. R. (1998) 'Formative assessment: revisiting the territory', *Assessment in Education: Principles, Policy & Practice*, vol. 5, no. 1, pp. 77–84.
Scouler, K. M. & Prosser, M. (1994) 'Students' experiences in studying for multiple choice question examinations', *Studies in Higher Education*, vol. 19, no. 3, pp. 267–279.
Sharpe, R., *et al*. (2006) *The Undergraduate Experience of Blended -Learning: A Review of UK Literature and Practice*, T.H.E. Academy, New York, UK.
Trowler, V. & Trowler, P. (2010) 'Student engagement evidence summary', *Higher Education Academy*, vol. 2010, pp. 1–34.
Turney, C. S. M., *et al*. (2009) 'Using technology to direct learning in higher education', *Active Learning in Higher Education*, vol. 10, no. 1, pp. 71–83.
Voelkel, S. & Bennett, D. (2013) 'New uses for a familiar technology: introducing mobile phone polling in large classes', *Innovations in Education and Teaching International*, [online] Available at: http://dx.doi.org/10.1080/14703297.2013.770267