The impact on learning outcomes in mathematics of mobile-enhanced, combined formative and summative assessment

Knut Bjørkli

Faculty of Technology,
Sør-Trøndelag University College,
E.C. Dahls Gate 2, 7004 Trondheim, Norway
Email: knut.bjorkli@hist.no

Abstract: This paper presents the results of a study done on an introductory mathematics course for engineering students at a university college in Norway, in which a web-based assessment and classroom response system called Peer Learning Assessment System (PeLe) was used to run multiple choice tests encompassing both formative and summative elements. Using a between-subjects design and a mixed method approach, we have conducted both qualitative studies of students' attitudes towards combined formative and summative assessments, as well as quantitative studies of the effect of mobile-enhanced learning on learning outcomes. Students perceive the combination of formative and summative assessment as highly beneficial to their learning and they identify key factors for how both formative and summative aspects of assessments should be carried out to maximise learning. We also present quantitative data that support the conclusion that mobile-enhanced assessment can improve learning outcomes, with effect sizes around 0.5.

Keywords: formative assessment; summative assessment; mobile learning; impact; mobile-enhanced; learning outcomes.

Reference to this paper should be made as follows: Bjørkli, K. (2014) 'The impact on learning outcomes in mathematics of mobile-enhanced, combined formative and summative assessment', *Int. J. Technology Enhanced Learning*, Vol. 6, No.4, pp.343–360.

Biographical notes: Knut Bjørkli is an Assistant Professor at the Faculty of Technology in Sør-Trøndelag University College. He teaches undergraduate courses of mathematics and physics. Since 2009 he has worked on numerous international projects focusing on mobile-enhanced learning, particularly the use of classroom response systems and digital assessment. His current research interest is methodologies for maximising learning impact of technology-enhanced formative assessment.

1 Introduction

The distinction between formative and summative assessment in the context of student assessment was first made by Bloom (1969). Bloom defined formative assessment as an evaluation whose purpose is '… to provide feedback and correctives at each stage in the teaching-learning process', whereas the purpose of summative assessment is to measure what a student has learnt at the end of a course.

Exactly what constitutes formative assessment has become a subject of some discourse (Cech, 2007; Bennett, 2011). On one side of the argument, there are those who define formative assessment as a diagnostic, interim test taken at some point during a course, in which data is collected and interpreted to shape the learning process. Those on the opposite side of the debate see formative assessment as a process rather than a set of tests, arguing that 'Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes' (McManus, 2008, p.3).

For the purposes of this paper, we will adopt a definition of formative assessment which incorporates both of the above views: formative assessments are interim tests taken over the course of a term, which are part of a process used by both teacher and students to improve learning outcomes. The tests are used to diagnose students' problem areas and then the test data are used to provide feedback to students on their performance and problem areas. Test data also 'feed forward' to shape the learning process which immediately follows the test, as well as aid in structuring the traditional lectures which follow.

Researchers report varying effectiveness of formative assessment as to impact on learning outcomes. An often used metric is that of effect size, which in the context of learning outcomes is defined as the difference in average scores between a test group and a control group, divided by the standard deviation of the test group. The seminal works of Black and William (1998) is often used by advocates of formative assessment to demonstrate large effect sizes, some in the range of 0.4–0.7 (Stiggins, 1999; Popham, 2008, p.19; both referenced in Bennett, 2011) all the way up to effect sizes of one to two standard deviations (Bloom, 1984; Stiggins, 2006, p.15).

Other sources report more modest effectiveness of formative assessment. A comprehensive meta-study by Kluger and DeNisi (1996) found average effect sizes of around 0.4 with regards to the effect of providing feedback on performance and noted that '(…) researchers have recently recognised that FIs [feedback interventions] have highly variable performance, such that in some conditions FIs improve performance, in other FIs have no apparent effects on performance and yet in others FIs debilitate performance…' (Kluger and DeNisi, 1996, p.1).

The main purpose of summative assessment is, by definition, to provide some measure of the degree to which learning objectives have been met and to a lesser extent, to help shape the learning process. However, there is ample evidence that properly implemented summative assessments can aid learning (Zwick et al., 2001; Shepard, 2006; Popham, 2008, p.23; Corcoran et al., 2009; Rohrer and Pashler, 2010; all referenced in Bennett, 2011).

Mobile-enhanced learning tools can help with harnessing the learning potential of both formative and summative assessments: such tools can facilitate the learning processes that emerge as a result of formative assessments and also reduce the workload on the teacher for summative assessments, by automatically correcting and grading tests. This paper presents a study of the use of one such mobile-enhanced learning tool called Peer Learning Assessment System (PeLe), for combined formative and summative assessments in the subject of mathematics.

First we present the context in which PeLe was developed and why it was chosen over existing technologies, before outlining the research objectives and the mixed-method research procedure (Jick, 1979) used. We present both quantitative results from the summative assessments in terms of impact on learning outcomes, as well as

qualitative results from student surveys and interviews focusing on attitudes towards combined formative and summative assessments, before a brief discussion of study limitations and a conclusion.

2 Context

2.1 The peer learning assessment system (PeLe)

Sør-Trøndelag University College (HiST) has since 2009 been devoting a significant portion of its research and development efforts to mobile-enhanced learning and assessment, as part of a long-term mobile learning strategy. By the time of writing, the research has resulted in a portfolio of technologies and methodologies consisting of three separate mobile technologies which address various aspects of the learning and assessment process: an online student response system called SRS (EduMecca, 2010); a system for student evaluation called Eval (iQVET, 2014) and the assessment system PeLe (Done-IT, 2011), with the latter being the focus of this paper.

There exists a multitude of tools which can be used for combined formative and summative assessments – e.g., learning management systems like its learning, moodle, blackboard, etc. However, most existing tools focus on asynchronous, self-paced, student-controlled assessment, as opposed to synchronous, teacher-controlled assessments that take place in the classroom. Furthermore, most existing technologies are designed to run on computers and can be cumbersome to navigate on mobile devices.

PeLe is an online assessment tool for multiple-choice tests, designed to be used with students' mobile devices and consists of two main components: (a) the teacher client, which is installed on the teacher's PC and is used during classes to run and monitor assessments and also to structure learning activities, such as teacher review of questions and group discussions which follow immediately after the test; (b) the student client, which is a web-based interface used by the students to answer assessments from a mobile device. These two components are shown in Figure 1.

Figure 1 The PeLe teacher client showing the performance of the class on each test question (left) and student client showing students in the process of answering the test using a mobile device (right) (see online version for colours)

The teacher sets up the test beforehand, by setting the number of questions for the test; the number of alternatives on each question and by selecting which alternatives are the correct ones. PeLe will then automatically grade the test once all the students have submitted their answers – but the results will not be made available to the students until the learning activities immediately following the test have been completed.

The continuously updating result graphs help the teacher to quickly identify problem areas which need to be addressed and even as the test is in progress, the teacher will start planning learning activities for the review phase which starts following a short break for the students. PeLe has a facility by which students can 'flag' questions in order to signal uncertainty about the answer and to highlight the need for a particularly thorough review.

The teacher uses the break to refine the structure of the learning activities that will take place during the review phase, using in particular the distribution of correct/incorrect answers as a guide. For example: the difficult items which the majority of students answered incorrectly may require the teacher to explicitly demonstrate a correct solution strategy. This approach doesn't involve peer learning, but it does give students immediate feedback on their reasoning during the test.

Making mistakes is a crucial part of any learning process and PeLe has a facility to give students a second chance to answer a test question during the review phase – after a short discussion in which the students work in small groups to identify errors in each other's reasoning. This peer learning approach was used for items which were correctly answered by a significant percentage of students (50–70%) and enabled students to demonstrate that they were able to learn from their mistakes. A student who answered a test question incorrectly the first time around, but got it right during the review phase, received partial credit for that.

Once the test is completed and all post-assessment learning activities are finished, the results are made available to the students at the teacher's discretion – students can then log on to the student interface to view their performance on each question, as well as an overall score.

3 Research objectives

Mathematics is a core subject in engineering education and is considered a difficult subject for students to grasp – see Shaw (1999) for a review of student attitudes towards the subject of mathematics at universities in the UK. Given the prominent position of the subject in the curriculum, an introductory mathematics course was a suitable test bed for a study of the influence of formative and summative assessment on learning outcomes. The impact on learning outcomes of either formative or summative assessments has been studied quite extensively, but the combination of formative and summative assessment has mostly been studied in the context of large-scale, national assessments – see for example Looney (2011).

The study presented in this paper approaches the subject of combining formative and summative assessment on a smaller scale and we formulated two main research questions for this study:

1 How does mobile-enhanced, combined formative and summative assessment influence learning outcomes? Specifically: do students who follow a mathematics course using such an assessment approach significantly better than a control group of students who follow the same course without combined formative and summative assessments?

2 What are the students' attitudes towards combining formative and summative assessments in the way that was done for the mathematics course? In particular, how do the students weigh the importance of the formative aspects (using tests to shape the way students learn) and the summative aspects (performance and grading), with respects to the impact on learning outcomes?

The fact that there were three large classes (n > 100 in each class) following this particular mathematics course, allowed a between-subjects design for this study – i.e. each student is subject to either the test treatment, or the control treatment. This eliminates carryover effects associated with within-subjects studies, in which every student is subjected to all treatments – both test and control treatments.

4 Procedure

The study was conducted on a 15 week, 10 ECTS-credits introductory mathematics course for chemical and logistics engineering students (n = 113) at Sør-Trøndelag University College (HiST) in Trondheim, Norway during the autumn term of 2013. Two control groups at HiST following the same curriculum were selected:

- Control group 1: a class of building engineering students (n = 112)
- Control group 2: a class of mechanical engineering students (n = 96).

All groups had three 90 minutes lectures per week, but for the test group, a total of six lectures were converted into assessment sessions using PeLe, at intervals of two–three weeks. So in effect, the test group had six;·1.5 hours = 12 hours less of traditional lecturing than the control groups.

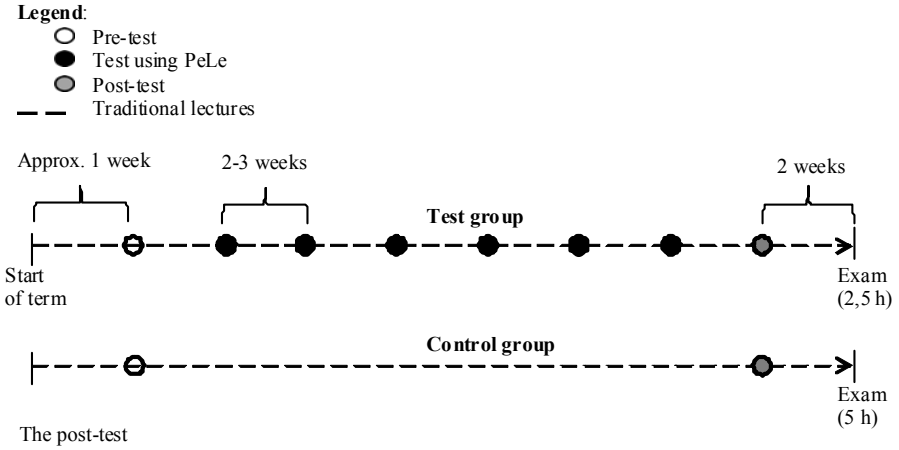
A standardised pre-test created by the Norwegian Mathematical Council was administered at the very beginning of the course, designed to measure proficiency in key mathematical concepts like basic arithmetic, algebra, percentages and geometry. The test was a combination of constructed-response items and multiple-choice questions which was answered on paper and the purpose of the pre-test was to uncover any differences in the level of mathematical ability that might exist between the groups. The baseline provided by the pre-test would then be compared to results from a post-test set two weeks before the final exam, in addition to the final exam itself, to measure any shift in relative learning outcomes between the test group and the control group.

A timeline of the study is illustrated in Figure 2.

The post-test was a multiple-choice test which contained questions from all parts of the curriculum and both the test group and the control group answered identical versions using mobile devices, to augment the quantitative data provided by the final exam. The post-test was set as late in the term as practically possible, approx. two weeks before the final exam, to enable the post-test to incorporate as much of the course curriculum as possible.

For both the test and control group, the final exam was a traditional, written exam with students working individually under supervision and which did not use mobile technology at any level.

Figure 2 Timeline of the study for the test group (top) and control group (bottom)



For the control group, the final exam contained 24 questions and counted 100% of the final mathematics grade. The duration of the exam was five hours.

The test group's final exam had 12 questions, containing a representative subset of problems from the exam paper of the control group. Because the six tests taken during the term counted 60% of the final grade, the duration of the final exam, counting the remaining 40%, had to be reduced to 2.5 hours; half the duration of the control group.

4.1 The test group

The main method of instruction for the test group was traditional 90 minutes lectures, with three lectures per week. The lectures took the form of teacher writing notes on a whiteboard, interspersed with some interactive problem-solving sessions where the students were given mathematics problems to work with, and followed by a review by the teacher. No mobile technology was used during these lectures.

A total of six lectures, which were set approximately two–three weeks apart, were converted into sessions dedicated to multiple-choice assessments, replacing the traditional lecture for that day. Each assessment session lasted 100 minutes spread across two 45 minutes lectures, including a ten minutes break. The tests were handed out to the students on paper at the beginning of the class.

Each test contained five–six questions, with test questions being made up of standard mathematical assignments, encompassing both conceptual and computational exercises. At the beginning of the assessment, the test items were presented in a constructed-response format – i.e. the questions were initially not posed as multiple-choice. This was done to stimulate a deductive mathematical approach to problems and to avoid that students use the alternatives to 'reverse engineer' the correct answer, e.g., by method of elimination.

The tests were not accumulative, in the sense that each test introduced new material and did not rely directly on the curriculum of previous tests.

After about 20 minutes, the students were handed the multiple-choice version of the test and the students then collated their constructed responses to the alternatives given and answered the test using PeLe. The teacher monitored the students' answers in real time, to identify problem areas and prepare corrective actions for the review phase.

An example of a test question is given in Table 1, showing both the constructed-response item, as well as the multiple-choice version of the question.

Table 1 Example of test question from the subjects of population growth, differential equations and initial value problems, with both the constructed-response item given to students at the beginning of the test (left) and the multiple-choice version used to answer the question with PeLe (right)

Constructed-response item	*Multiple-choice version*
A colony of banana flies has a population size of 100 at $t = 0$. The population grows at a rate proportional to the population size at any given time.	A colony of banana flies has a population size of 100 at $t = 0$. The population grows at a rate proportional to the population size $y(t)$ at any given time.
Letting y denote the population size as a function of time, set up an initial value problem for $y(t)$.	Which initial value problem correctly describes $y(t)$? A: $\frac{dy}{dt} = 100t$ B: $\frac{dy}{dt} = 100 + ky(t)$ C: $\frac{dy}{dt} = y(t), \quad y(0) = 100$ D: $\frac{dy}{dt} = k \cdot y(t), \quad y(0) = 100$ E: $\frac{dy}{dt} = 100, \quad y(0) = 0$

Once the test was finished there was a short break, followed by the review phase. By this time, the teacher had complete overview of which questions were poorly answered and could devote time to the difficult ones.

Depending on the percentages of correct/incorrect answers, different learning activities would be initiated by the teacher: For questions with a high percentage of correct answers, the teacher would provide a brief verification of what the correct answer was and why.

For questions which had a percentage of correct responses of around 30–70%, the teacher would initiate peer learning activities like group discussions, after which students were given a chance to participate in 'second chance' voting sessions using PeLe. Answering a test question correctly in a 'second chance' session during the review phase would give partial credit, even if the student answered the test question incorrectly the first time around. This PeLe mechanism was designed to let students demonstrate the ability to learn from their mistakes – i.e. to realise during group discussions that their original reasoning had been incorrect and to then correct themselves a second time around.

The students' participation in peer learning activities was not subject to summative assessment – i.e. the level of participation or performance during discussions did not count towards the test score. Test scores were always awarded on an individual basis based on the ability to answers test questions correctly – either the first time around, or during second-chance sessions.

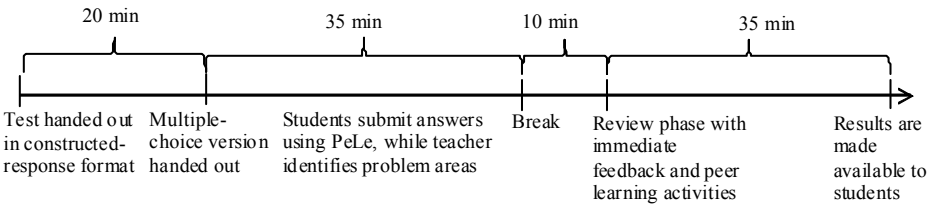
On some occasions, a large proportion of students had answered a particular question correctly, but many students had 'flagged' the question to indicate that they were unsure about their answers. The teacher could then pose a follow-up question during the review phase to gauge the depth of the students' understanding of a particular concept. A follow-up question used one of the test questions as a starting point, but either focused on a slightly different aspect of a problem, or introduced a slight modification of the original problem. The follow-up question was then answered as a multiple-choice question using PeLe, after the students had been given some time to work on the question either individually or in groups.

It was up to the teacher to decide at what point the students would see the result graphs and to reveal what the correct answer was. Normal procedure was to let the students see the result graphs before the review of a particular question (to give a collective sense of how easy/difficult the question was) and not reveal the correct answer until learning activities had been completed.

Once the learning activity pertaining to a particular question had been completed, the correct answer was revealed by the teacher. At the end of the assessment session, individual scores and grades were made available to the students.

The timeline of an assessment session using PeLe is illustrated in Figure 3.

Figure 3 Timeline of an assessment session using PeLe



The mathematics course, in which this study was made, was the only instance of combined formative and summative assessment that the test group was subjected to.

4.2 The control groups

The two classes comprising the control groups were taught by two separate teachers, none of whom taught the test group. Like the test group, the control groups had three 90 minutes classes per week, which were delivered by traditional classroom lectures. Some interactive elements were incorporated into the lectures, such as students solving problems in class and the teacher providing a review of those problems, but no form of mobile technology was used during lectures.

The control groups did not use formative assessments of any kind during the term, in any of the subjects taught.

5 Datasets

5.1 Measuring the effect of mobile-enhanced formative assessment on learning outcomes

To measure any shift in learning outcomes between the test group and the control groups, a dataset comprising the results of the three comparative tests shown in Figure 2 was defined: a pre-test, in the form of a standardised test by the Norwegian mathematical council, taken at the beginning of the term; a post-test in the form of a multiple-choice test comprising most of the course curriculum, taken two weeks before the final exam and answered using mobile devices; and the final written exam. The scores of the test group and control groups on each test were then compared and checked for statistically significant differences.

5.2 Examining students' attitudes towards combining formative and summative assessments

To gain insight into the students' attitudes towards integrating both formative and summative elements into assessments, a qualitative approach was used: an online survey was conducted at the middle of the term (shortly after the second PeLe test in Figure 2), in addition to two focus group interviews. One interview was conducted shortly after the first test, while the second interview was conducted approximately one month after the final exam, after the term had ended.

6 Results

6.1 Results from summative assessments

Scores from assessments and the final exam were graded using the following grading scale.

Table 2 Grading scale used during assessments and exams

A	90–100%
B	75–90%
C	60–75%
D	45–60%
E	35–45%
F – fail	0–35%

Table 3 summarises the results of the pre-test, the post-test and the final exam, respectively, for the test group and the two control groups, using the grading scale in Table 2. Here n is the number of students participating in the test; M indicates the average score for the group in terms of percentage of the total score and SD is the standard deviation of scores for each group.

Table 3 Results of tests taken by all the groups during the term

Group	*Pre-test*			*Post-test*			*Exam*		
	n	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Test group	100	51%	18%	59	64%	15%	113	64%	18%
Control group 1	94	54%	19%	61	43%	17%	100	54%	23%
Control group 2	78	54%	19%				80	52%	24%

For the post-test, the two control groups were merged into one single control group to achieve statistical significance, since a limited number of students from each of the two control groups participated.

To check whether differences in scores between the test group and the control groups from Table 3 were statistically significant, two metrics were used: (a) a statistical hypothesis test using a two-sample student *t*-test was performed as described by Marx and Larsen (1986) and (b) calculation of the effect size in terms of the parameter Cohen's *d* (Cohen, 1988).

6.2 Hypothesis testing using the two-sample student t-test

Firstly, a two-sample *t*-test was performed, using the following null hypothesis:

H0: The average score of the test group is equal or less to that of the control group

This was tested against the alternative hypothesis:

H1: The average score of the test group was larger than the score of the control groups

The hypothesis test was then performed at a 5% level of significance on the results of the pre-test; the post-test and the final exam.

To decide whether to reject the null hypothesis, the *t*-test calculates the test statistic *t*, which is then compared to the critical value $t_{critical}$. The null hypothesis will be rejected if $t > t_{critical}$ at the given level of significance.

Table 4 Student *t*-test performed on the test group versus the two control groups at 5% level of significance

	Pre-test			*Post-test*			*Final exam*		
	t	$t_{critical}$	*Reject H0?*	*t*	$t_{critical}$	*Reject H0?*	*t*	$t_{critical}$	*Reject H0?*
Test group vs. control group 1	1.15	1.97	No	7.28	1.98	Yes	3.45	1.97	Yes
Test group vs. control group 2	1.12	1.97	No				3.98	1.97	Yes

For the post-test, the two control groups were merged into one single control group since a limited number of students from each of the two control groups participated. The calculated t statistic for the post-test was approximately identical in both control groups and so the merging operation did not affect the conclusion to reject the null hypothesis.

The entries of Table 3 show that the following inferences hold, at the given level of significance: For the pre-test, even though the test group attained a lower score than the control groups, the difference is not statistically significant. Therefore, at the start of the term, the level of academic performance of the test group can be considered equal to that of the control groups. This pattern had shifted for the post-test, as the test group scored significantly better than the control groups. The same was observed for the final exam, as the test group scored significantly better than the control groups.

It should be noted that the post-test data is less reliable than the data from the pre-test and the exam, since nearly half the students in the control group did not participate in this particular test. It is therefore likely some of the higher-performing students were absent and that this skewed the results for the control group. The data from the post-test will therefore not be used to compare learning outcomes between the test group and the control group.

6.3 Calculation of effect sizes in terms of Cohen's d

To augment the hypothesis tests and to provide more detailed information about the extent to which scores differed between the test and control groups, the effect size in terms of Cohen's *d* was calculated.

If $\overline{x}_{test}$ and $\overline{x}_{control}$ denote the calculated average scores (means) of the test and control group, respectively, and s is the pooled standard deviation of the groups' scores, then Cohen's d is calculated as:

$$d = \frac{\overline{x}_{test} - \overline{x}_{control}}{s}$$

We will adopt the convention of (Cohen, 1988, p.25) to define the effect size in terms of Cohen's *d*: the effect size is 'small' when $d = 0.2$; 'medium' when $d = 0.5$ and 'large' when $d = 0.8$. We will also follow the standard convention that effect sizes are positive when it is in the direction of improvement (i.e. higher average scores for the test group).

The data in Table 5 are in compliance with Table 3: For the pre-test, the negative effect size reflects a lower score for the test group compared to the control groups, but the absolute value is lower than Cohen's threshold for 'small' and indicates that the test group can be considered equal to the control groups.

Table 5 Calculation of effect sizes in terms of Cohen's *d* when comparing average scores between test group and control groups

	Pre-test	*Post-test*	*Exam*
Test group vs. control group 1	–0.163	1.320	0.493
Test group vs. control group 2	–0.164		0.490

For the post-test, the effect size exceeds 'large' and supports the conclusion that the test group scored significantly higher than the control groups. For the final exam, effect size is around $d = 0.5$, which corresponds to 'medium'. It indicates that while the test group scored significantly higher than the control groups, the effect is not as pronounced as the post-test.

Since the main differentiating factor between the test group and the control groups was that the former was subjected to mobile-enhanced, combined formative and summative assessments, the data indicates that combined formative summative assessment, in the way it was done in this study, can improve learning outcomes.

The 'medium' effect size of around 0.5 found in this study for the final exam is in agreement with much of the research literature on the impact of formative assessment on learning outcomes.

6.4 Distribution of grades on final exam

We have already demonstrated positive effect sizes on learning outcomes by using mobile-enhanced, combined formative and summative assessments, but we would like to look in some detail at what type of students benefit the most from this approach – whether it's the lower-performing student, or the high achievers.

To that end, we have compiled the distribution of grades for the two groups, as shown in Figure 4 (the grading scale of Table 2 is used).

Figure 4 Distribution of grades for final exam of test group and control group (see online version for colours)

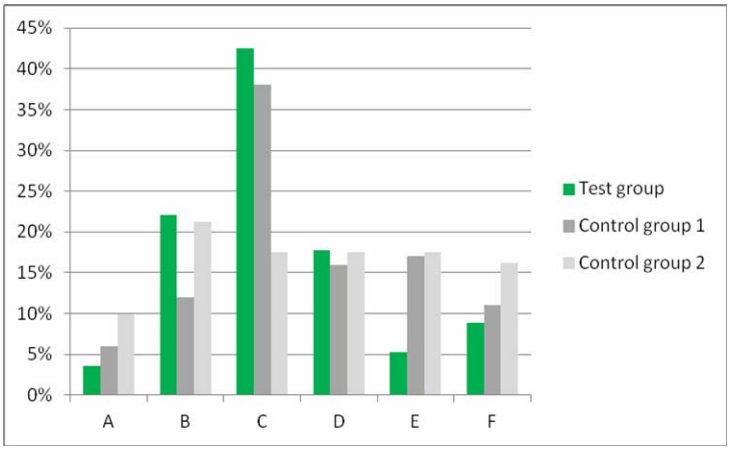
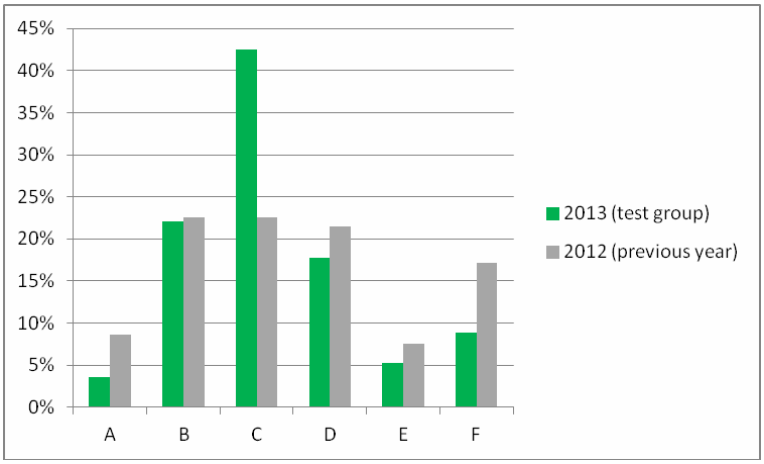


Figure 4 indicates that the number of students scoring E (lowest passing grade) or F (failing grade) is reduced in the test group, relative to the two control groups. There is a clustering of students around the grade of C, corresponding to an 'average' exam performance.

In an attempt to isolate the effects of mobile-enhanced combined formative and summative assessments and the effects of the teacher, respectively, it is illustrative to compare the result of the study presented in this paper, to the results of the previous year. The same teacher, who taught the test group of chemical and logistical engineering students in this study, also taught the class of chemical and logistical engineering the previous year – during which no combined formative/summative assessment activities took place. The comparison is illustrated in Figure 5.

Figure 5 shows the same pattern as Figure 4, in that the test group has fewer E's and F's and a higher percentage of C's than the previous year's class. Both datasets indicate that it is the lower-performing students who receive the biggest gains in performance from the methodology presented in this study.

Figure 5 Comparison of exam grades for the class of mathematics for chemical and logistical engineering students for the past two years (see online version for colours)



7 Discussion

In line with a mixed method approach, the quantitative impact on learning outcomes outlined in the previous section was augmented by a qualitative study of the test group students' attitudes towards combined formative and summative assessment. To what extent did they perceive this approach as beneficial to learning and which of the two aspects of the assessments – formative or summative – did they feel have the biggest impact on learning outcomes? And how do the students' observations compare to the quantitative data on learning outcomes?

7.1 Results from the survey

An online survey was run for the test group about one month into the term (shortly after the second PeLe test in Figure 2) and was used to gain insight into students' attitudes towards combined formative and summative assessments in the manner that was performed in this study.

Firstly, students were asked what their primary motivation was for studying mathematics. Their answers are summarised in Figure 6.

When asked about their attitude towards combined formative and summative assessments using PeLe, the response was as shown in Figure 7.

Figure 6 Students' primary motivational factor in the subject of mathematics (see online version for colours)

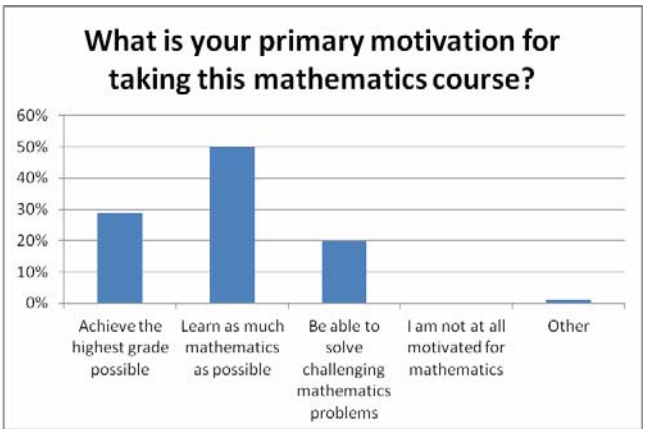
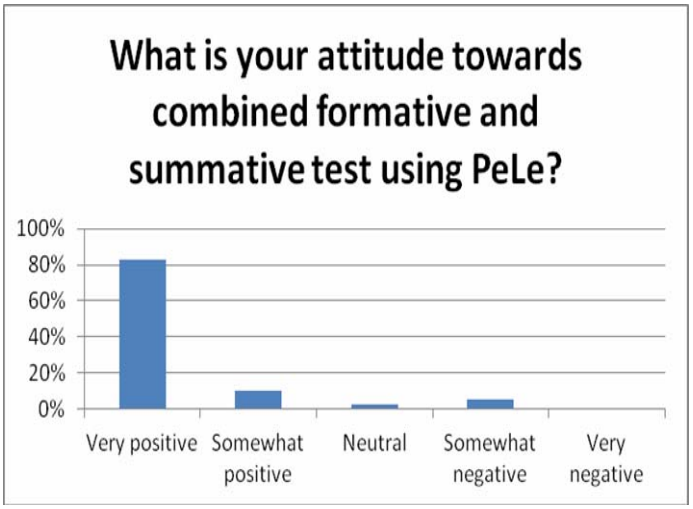


Figure 7 Students' attitudes towards combined formative and summative assessments with PeLe (see online version for colours)



When asked to elaborate, students reported that the tests stimulated to continuous works and that it reduced pressure and stress about the final exam.

On the negative side, some students took exception to limitations of multiple choice tests. They felt that multiple choice questions in mathematics are rigid and inflexible, since such questions do not allow them to show logic, reasoning and detailed calculations.

The students also reported that there was limited space available in the classrooms used to conduct the tests and students were sitting so tightly packed that it became tempting to cheat by looking at other students' answers.

When asked about the perceived effect on learning outcomes of the assessments, 90% of the students reported a high or fairly high learning effect of such assessments.

7.2 Results from the interviews

Two focus-group interviews were conducted: the first one was done four weeks into the term, with four students participating (two male and two female). The same students then participated in a follow-up interview one month after the final exam, by which time the term had ended and exam grades had been made available to the students. Students volunteered to participate in these focus groups and participants were selected randomly.

One section of the interviews focused on the formative aspects of the assessments and the students identified several key factors for how the formative assessments should be conducted in order to maximise the perceived impact on learning outcomes.

Group discussions were not seen as particularly enlightening, unless there was a good match in academic ability between the students discussing. Also, the students reported that group discussion did not allocate sufficient time for students to reason together and the end result was often, in the words of one of the students, '(…) less able students allied themselves with high-performing ones and chose whichever alternatives they outlines as being correct'.

During the review phase, the students preferred answering follow-up questions that were similar, but not identical, to the test questions. 'This let you demonstrate that you have fully grasped a concept', as one student put it. To be given a second chance to answer a question from the test (e.g., following a group discussion of a question where a significant portion of students got it wrong the first time) was perceived as less beneficial for learning – compared to answering a similar, but different question.

Handing out test questions in a constructed-response format at the beginning of the test was considered beneficial for learning, as this forced the students to approach the mathematical problems in a traditional, deductive manner – as opposed to the guesswork and method of elimination to rule out the least plausible answers that would often ensue if they were given the multiple-choice version at the outset.

The students were asked at what time they felt it appropriate to be given their test scores. Some students argued that scores should be made available directly after the test had finished – i.e. the results should be made available to the students before the review phase. Students who argued this position felt that the excitement or insecurity over their test scores which resulted from withholding scores until the end of the review phase, detracted attention from the learning activities. However, these students had not reflected on, or were not challenged on during the interview, the fact that revealing test scores ahead of the review phase would render some of the peer learning activities void (e.g., discussing test questions before 'second chance' voting sessions).

Due to a shortage of sufficiently large classrooms, the tests had to be held in auditoria used for lectures, in which the students sat tightly packed. The students reported that the lack of space made it tempting to cheat, by glancing at answers of nearby students.

It was felt that the tests, which counted 60% towards the final grade, were given too much weight compared to the final exam.

When the students were given a second chance to answer a test question, the student felt that the follow-up answer should carry less weight than the original answer.

During discussions, there was a tendency for students to ally themselves with other students they knew to be good at mathematics, in order to maximise their score during 'second chance' voting.

A key part of the interview was trying to elicit from the students which aspect of the assessment - formative vs. summative aspect – they felt had a greater impact on their

learning. All four students were unanimous about this: the fact that the tests counted towards a significant percentage of the overall grade was seen as profoundly important. For this reason, students worked conscientiously with the subject in front of upcoming tests. Conversely, if the assessments had been purely formatively, the students said they would not have put as much effort into neither the preparations for the tests, nor the test themselves.

At the time of the interviews, the students were not aware of how they – as the test group – had performed at the final exam compared to the control group. Given that the quantitative data showed positive effect sizes on learning outcomes in favour of the test group, the students' statements imply that both the formative and summative aspects of the assessments had an effect on learning outcomes: The learning activities performed directly after each test enhanced student learning, while the fact that tests counted towards the final grade was a major motivational factor for students to work continuously on the subject and to prepare thoroughly for the tests.

8 Limitations

In this section we discuss the main limitations of the study, which should be taken into consideration when analysing the data.

8.1 Differences in teacher performance

At the end of the term, all three teachers were rated by their respective students for their ability to stimulate and motivate students into working with the subject of mathematics; their ability to engage and communicate with students; and their ability to explain theory and concepts to students. Each of the three teachers was given close to maximum score on all points and differences in pedagogical skills between teachers are likely to be a very small effect in this study.

8.2 Differences in grading practices between teachers

The pre-test was graded using an unambiguous scoring system laid down by the Norwegian Mathematical Council. The post-test was graded automatically by the PeLe software and was therefore not subject to teacher grading.

The final exam of each group was corrected and graded by the group's teacher. To check whether there were systematic differences in teacher grading practices, the teacher of the test group swapped exam papers with the teachers of the control groups and scores were compared. No systematic differences in grading practices were identified.

8.3 Different exam content and duration

The test group had a final exam lasting 2.5 hours, while the two control groups had an exam lasting five hours. Due to different exam lengths, the test group's exam was not identical to that of control group – instead, the test group's exam contained a selected subset of problems from the control group's exam.

Three experienced mathematics teachers at the faculty were independently asked to rate the relative difficulty levels of the test group's 2.5 hours exam to the five hours exam of the control groups. All teachers concurred that the difficulty level of test group's exam was equal to, or higher than that of the control groups, since it included all the more difficult items while shedding the more trivial ones.

8.4 Separating effects of technology and formative assessment

The design of this study makes it impossible to separate the effects of using technology (i.e. using mobile devices to answer multiple-choice tests) from the effects of using formative assessments. Only by letting at least one of the control groups use mobile technology without the formative aspects would such a separation have been possible, but this was not done due to time constraints (there was not enough time or manpower available to train teachers in the control groups in the use of PeLe).

9 Conclusion

We have reported findings from a mixed-method study of the use of mobile-enhanced, combined formative and summative assessments in a mathematics course for first-year engineering students at a university college in Norway. In focus group interviews, students highlighted the summative aspect of the assessments as essential to learning outcomes, as the desire to post good grades made them work consistently with the subject throughout the term. Students also identified key factors for maximising the perceived learning effect of the formative aspects of assessments. Within the limitations of the study, we find that mobile-enhanced, combined formative and summative assessment have improved learning outcomes in the test group, with 'medium' effect sizes of around 0,5 for the final exam.

10 Future research

To address some of the limitations of this study – in particular, to de-couple the effects of mobile technology from the effects of using formative assessments – we will undertake a new study in which both test and control groups use mobile technology, but reserve the formative aspects for the test group. We will also conduct studies into the relative effects of formative and summative aspects, respectively, when doing combined formative and summative assessments with PeLe – to see how the two aspects influence student learning outcomes.

References

Bennett, R.E. (2011) 'Formative assessment: a critical review', *Assessment in Education: Principles, Policy & Practice*, Vol. 18, No. 1, pp.5–25.

Black, P. and William, D. (1998) 'Inside the black box: raising standards through classroom assessment', *Phi Delta Kappan*, Vol. 80, No. 2, pp.139–148.

Bloom, B.S. (1969) 'Some theoretical issues relating to educational evaluation', in Tyler, R.W. (Ed.): *Educational Evaluation: New Roles, New Means*, *The 63rd Yearbook of the National Society for the Study of Education, Part 2*, University of Chicago Press, Chicago, IL, USA, Vol. 69, pp.26–50.

Bloom, B.S. (1984) 'The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring' *Educational Researcher 13*, no. 6: 4–16.

Cech, S.J. (2007) 'Test industry split over 'formative' assessment', *Edweek*, Vol. 28, No. 4, pp.1–15.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed., Lawrence Earlbaum Associates, Hillsdale, NJ, USA.

Done-IT (2011) *Done-IT Project*. Available online at: http://histproject.no/node/167 (accessed on February 2014).

Edumecca (2010) *Edumecca Project*. Available online at: http://histproject.no/node/18 (accessed on February 2014).

iQVET (2014) *iQVET Project*. Available online at: http://histproject.no/node/167 (accessed on February 2014).

Jick, T.D. (1979) 'Mixing qualitative and quantitative methods: triangulation in action', *Administrative Science Quarterly*, Vol. 24, No. 4, pp. 602–611.

Kluger, A.N. and DeNisi, A. (1996) 'The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory', *Psychological Bulletin*, Vol. 119, No. 2, pp.254–284.

Looney, J.W. (2011) *Integrating Formative and Summative Assessment: Progress Toward a Seamless System?*, *OECD Education Working Papers*, No. 58, OECD Publishing.

Marx, M.L and Larsen, R. (1986) *An Introduction to Mathematical Statistics and its Applications*, 2nd ed., Prentice-Hall.

McManus, S. (2008) *attributes of Effective Formative Assessment*, Council for Chief State School Officers, Washington, DC, USA. Available online at: http://www.ccsso.org/publications/details.cfm?PublicationID=362 (accessed on 20 February 2014).

Popham, W.J. (2008) *transformative Assessment*, ASCD, Alexandria, VA, USA.

Rohrer, D. and Pashler, H. (2010) 'Recent research on human learning challenges conventional instructional strategies', *Educational Researcher*, Vol. 39, No. 5, pp.406–412.

Shaw, C.T. (1999) 'Attitudes of engineering students to mathematics a comparison across universities', *International Journal of Mathematical Education in Science and Technology*, Vol. 30, No. 1, pp.47–63.

Shepard, L.A. (2006) 'Classroom assessment', in Brennan, R.L. (ed.): *Educational Measurement*, 4th ed., American Council on Education/Praeger, Westport, CT, USA, pp.623–646.

Stiggins, R.J. (1999) 'Assessment, student confidence, and school success', *Phi Delta Kappan*, Vol. 81, No. 3, pp.191–198.

Stiggins, R.J. (2006) 'Assessment for learning: a key to motivation and achievement', *Edge*, Vol. 2, No. 2, pp.3–19.

Zwick, R., Senturk, D., Wang, J. and Loomis, S.C. (2001) 'An investigation of alternative methods for item mapping in the national assessment of educational progress', *Educational Measurement: Issues and Practice*, Vol. 20, No. 2, pp.15–25.