PeerWise: Students Sharing their Multiple Choice Questions

Paul Denny, John Hamer and
Andrew Luxton-Reilly
Dept. of Computer Science
University of Auckland
Auckland, New Zealand
{paul, j.hamer,
andrew}@cs.auckland.ac.nz

Helen Purchase
Dept. of Computing Science
University of Glasgow
Glasgow, United Kingdom
hcp@dcs.gla.ac.uk

ABSTRACT

PeerWise is a system in which students create multiple choice questions and answer those created by their peers. In this paper, we report on some quantitative results which suggest that students who use PeerWise actively perform better in final examinations than students who are not active. We note a significant correlation between performance in written (not just multiple choice) questions and PeerWise activity, suggesting that active use of the system may contribute to deep (and not just drill-and-practise) learning.

Categories and Subject Descriptors: K.3.1 Computers and Education: Computer Uses in Education

General Terms: Human factors.

Keywords: MCQ, peer assessment, automated, question test bank, PeerWise, contributing student.

1. INTRODUCTION

PeerWise is a web-based system that allows students to create multiple choice questions (MCQs) and answer those created by their peers. The tool supports student learning in a variety of ways. Students are asked to focus on the learning outcomes of a course by creating questions that align with these outcomes. Students improve their own understanding by writing an explanation of the answer to their question. After answering a question, the student can assign a quality rating and provide (anonymous) written feedback to the author. The author can also write a response to such feedback.

The act of evaluating the quality of a question and explanation engages students with the material and requires application of higher-order cognitive skills. Finally, the questions themselves provide a resource that students can use in traditional drill-and-practice form.

In this paper we compare the performance of students who are more actively engaged with PeerWise with those who are less engaged. We find that students who make greater use of PeerWise score significantly higher in final exams than students of equivalent ability who make less use of PeerWise.

2. RELATED WORK

The design of the PeerWise system has been reported in [7], and an analysis of the ways in which students use PeerWise across a range of courses can be found in [8]. Overall participation was found to be fairly uniform across students of different abilities, with students at all ability levels contributing more than the required number of questions and answers. Extensive voluntary use of PeerWise for drill and practice revision continues right up to, and in some cases after, the final exam.

PeerWise incorporates both self- and peer-assessment activities. Students use PeerWise to engage in elements of self-assessment by answering MCQs in a drill-and-practice fashion, and peer-assessment is facilitated through the question feedback forums. The literature on self-assessment [4, 10] and peer-assessment [17] reports a wide range of benefits, including:

- help to consolidate, reinforce and deepen understanding, by engaging students in cognitively demanding tasks: reviewing, summarising, clarifying, giving feedback, diagnosing misconceptions, identifying missing knowledge, and considering deviations from the ideal;
- highlight the importance of presenting work in a clear and logical fashion;
- expose students to a variety of styles, techniques, ideas and abilities, in a spectrum of quality from mistakes to exemplars;
- provide feedback swiftly and in quantity. Feedback is associated with more effective learning in a variety of settings. Even if the quality of feedback is lower than from professional staff, its immediacy, frequency and volume may compensate;
- promote social and professional skills;
- improve understanding and self-confidence; and

ICER'08, September 6–7, 2008, Sydney, Australia.
Copyright 2008 ACM 978-1-60558-216-0/08/09 ...$5.00.

- encourage reflection on course objectives and the purpose of the assessment task.

The nature of engagement with PeerWise encourages the development of an on-line learning community [18]. PeerWise can be used as part of a contributing-student pedagogy [6], in which students actively contribute to the learning resources available to the entire class. PeerWise can be used as one element of course assessment in courses designed to engender interaction and collaboration, although there are indications that it is perceived to be of limited value by students already involved in contribution-based activities in small classes [13].

Although concerns over the use of MCQs [3, 16] have been reported and they have a number of recognised limitations [14], their use for self-assessment purposes has been correlated with improved exam performance in [5]. Nicol [14] suggests that MCQs can be used to promote effective learning and satisfy the seven principles of good feedback reported in [15], for example, when students develop the MCQs themselves.

A number of systems related to PeerWise have been described in the literature.

Horgen used a lecture management system to share student generated MCQs in a class of 30 students [12]. Students worked in groups to create questions, and the activity included a required reflection stage. Horgen analysed the depth of the questions, and found that "only a few students actually managed to climb Bloom's ladder in their produced tests." This is consistent with our experience using PeerWise. Curiously, Horgen's students did not tend to use the question bank for drill-and-test. We surmise that this may be a consequence of a small class size, and hence small question bank.

Fellenz [11] reported on a course where students generated MCQs which were reviewed by their peers. Fellenz did not use technology to support the process. Students involved in the activity reported that the development of the MCQs helped develop a deep understanding of material because it required that they "made explicit their understanding of the complexities of the subject matter" [ibid, p.711]. Fellenz also reported that the activity increased student ownership of the material and motived students to participate.

Arthur [1] reports on a large course activity, in which the class is divided into streams of about 50 students. Each week, a small group of students from each stream work together to create a set of questions sufficient for a ten-minute quiz. These questions are submitted electronically, and the test is taken by students in another stream.

Yu [20] has students construct an MCQ item and submit the question to an on-line database. A peer-assessment phase is conducted where students provide feedback about the quality of the item and suggest improvements. After the evaluation phase, the questions are transferred to a test bank database to be used for drill-and-practice exercises.

Barak [2] reports on a postgraduate MBA course in which students contribute questions to an on-line repository. The students also rank their peers' contributions.

Chang [9] operate in a "one-on-one" educational computing classroom and have elaborated a theme of "asking a good question" in which each student generates a question and answer, then applies a self-assessment rubric before sending the question for peer-assessment. The mutual reviewers then form triads, each selecting two items for a class-wide discussion, during which the teacher points out any misconceptions and misunderstandings.

While all these reports agree that student-contributed MCQs is a powerful idea, this paper is the first to provide significant quantitative evidence supporting the approach.

3. DATA COLLECTION AND DEFINITION

In 2007, we studied the effects of PeerWise in a standard first-year programming course (CS101). The course is taken by both majors and non-majors in Computer Science. A test, which included MCQs, is held approximately half-way through the course. However, PeerWise was only introduced *after* this mid-semester test. Students were required to use PeerWise throughout the second half of the course, and the system remained available for voluntarily use in the study period leading up to the final exam.

Students were required to use PeerWise to create a minimum of two questions and to answer a minimum of ten questions. There was no requirement to provide any written feedback on the questions answered, although many students did so.

The analyses we performed require both the test and exam results of students in the course. For this reason, we have excluded data from students who either did not sit the test or did not sit the exam. The CS101 course had an original enrolment of 536. Of these, 48 did not sit the test and 70 did not sit the final exam. There were 76 students who had missed either the test or the exam, which left 460 students for the analysis.

We know that the characteristics of the top students in the class are very different than the weakest students [19]. In order to understand how the use of PeerWise affected different students, we divided the class into quartiles and asked whether the use of PeerWise provided any measurable benefit for the most capable students, average students and weak students. The quartiles were formed using the mark obtained in the mid-semester test (before any use of PeerWise), each quartile consisting of 115 students. Within each quartile, students were ranked on their level of "activity" with PeerWise, and we compared the most active half with the half that was least active.

3.1 Defining "activity"

Students use PeerWise in a number of ways, and any "activity" metric needs to take these different usages into account. The primary activities are: contributing new questions; answering existing questions; and writing open-ended comments on questions that have been answered. These give us three simple measures of activity: the number of questions contributed, the number of questions answered, and the number of comments that have been written.

The number of comments that a student has written provides some measure of their voluntary participation with PeerWise, as writing comments was not a required activity. However, it is still a fairly crude measure, as the actual content of the comments can vary greatly. There are many short comments that do not exhibit deep analysis of a question, such as "Nice", "Good one!", "excellent question". In contrast, there are also many insightful comments that demonstrate a critical and thorough analysis has been performed. Some examples are given in Table 1, and the degree of voluntary participation is summarised in Table 2.

While we are not able to look at all of the comments and

“Overall, you have put a lot of effort into this question, it is just a pity there seems to be a problem with it. Maybe you didn’t count from 0 when you were working out the alternatives? I calculated the first parameter as 11 (the index position of the 5) and so didn’t need to look at any of the distractors.”
“Nicely thought out question, I think it would have been better to call new String() again on the third line to be consistent with how you created the other Strings.”
“Good effort, but there are some problems with this code... for a start, it definitely will not compile. You cannot have a statement appearing *after* the return statement because that statement can never be reached, and the compiler will not allow this. Also, the actual question statement should be improved, to something like: ’When the program above is executed, which of the following outputs is not possible?’ because the word *false* has a particular meaning in Java. Apart from that, the idea for the question is very good, because you are assessing parameter passing as well as random numbers... good effort!”

“Without testing this, this is my thoughts:
String s1 = new String(”love”)
String s2=s1.substring(0,2); //s2 now equals ”lo”
String s3=new String(”is expensive”)
System.out.println(s2+s3)
by my rekoning, the output should be ”lois expensive”
By all accounts, your answer cannot be right anyhow, as your are printing out teh concatination of s2 and s3. s3 is ”is expensive”, so the output *MUST* have s2 followed by ”is expensive”. Since there is no space in s1, when you substring it, your string will be all letters, no whitespace, so when you concatenate them, there should NOT be a space inbetween s2 and s3, as your answer specifies.”

“My understanding is that it takes only ONE parameter, but that parameter itself contains an unlimited amount. For example you can’t have System.out.println(”Hello”, ”World”) but you can have System.out.println(”Hello” + ”World”) <note the comma>”

Table 1: Some of the more insightful student comments

	Total	Percentage
Number of students	460	
More than two questions	174	38%
More than ten answers	355	77%
One or more comments	326	71%

Table 2: Level of voluntary student participation (i.e. above the minimum requirements for assessment)

measure their quality (there are nearly 6000 comments for this course) we can use as a coarse measure for activity, the total character count of all comments a student has written.

For the CS101 course, the students had access to PeerWise from the 7^{th} May until the 14^{th} June 2007, a total of 39 days. Any day in which a student either contributes a new question, or answers an existing question is recorded as an “active” day. This count of the number of active days also provides another measure of student activity. We exclude days in which students log in to PeerWise and simply review questions they have written or questions they have already answered.

We define, for each student in the CS101 course, the following four individual measures of activity with PeerWise (and the abbreviations we have associated with each):

- total number of questions contributed (Q)
- total number of answers submitted (A)
- total character count of all comments written (C)
- number of days in which either a question was contributed or an answer was submitted (D)

We also defined one combined measure of activity (CM), as we felt this would more accurately describe a student’s overall level of engagement with PeerWise than any individual measure would. The combined measure was calculated by dividing each of the four measures above into deciles and summing the decile place for each student. For example, a student who ranked in the top 10% scores 10 for that metric, a student in the bottom 10% scores 1, etc. The scores for CM thus range between 4 and 40 (the actual minimum was 5; this is the only measure with a non-zero minimum).

Summary statistics for these measures are shown in Table 3.

Measure	1st Qt	Median	Mean	3rd Qt	Max
Q	2	2	2.6	3	28
A	11	19	34	36	600
C	0	184	707	772	16,410
D	1	3	3.3	4	26
CM	15	23	22	30	40

Table 3: Levels of activity, by category, for PeerWise use in CS101, for 460 students

4. COMPARISONS BETWEEN MOST AND LEAST ACTIVE STUDENTS

4.1 Methodology

We wished to determine how strongly engagement with PeerWise after the mid-semester test was linked to students’ achievement in the final examination. We did this by separating the 460 students in the class into achievement quartiles (with 115 students in each) based on their mid-semester test marks. Then, for each of the five different activity measurements defined above, we divided each quartile into two groups: those who were “Most PeerWise Active”(MPA) and those who were “Least PeerWise Active” (LPA) with respect to that measure.

Using a Student's t-test, we were able to determine, for each of the five activity measures, and for each quartile,

1. if there was any significant difference between the pre-PeerWise mid-semester test marks of MPA and LPA students (expecting this to not be the case);
2. if there was any significant difference between the post-PeerWise end of semester examination marks of MPA and LPA students (hoping that this would be the case, as it would confirm our hypothesis).

Each quartile contains 115 students, and, for each measure, each quartile was divided into two groups: LPA and MPA students. The division within each quartile was done by ranking the students according to the measure, and dividing them half-way. Students with the median score were randomly allocated to either the LPA or MPA groups. Thus there were 58 MPA and 57 LPA students in each quartile.

4.2 Hypotheses

We used a Student's t-test for independent samples with equal variance to compare the performance of the two groups. Our hypotheses are:

- H0: The mean examination mark of the MPA students is the same as the mean mark of the LPA students.
- H1: The mean examination mark of the MPA students is greater than the mean examination mark of the LPA students.

Note: this research hypothesis is directional (i.e. permits a one-tailed test of significance).

The test was held mid-semester, before use of PeerWise. It comprised 36 marks for MCQs and 64 marks for questions requiring written answers. The examination was held at the end of the semester, after use of PeerWise. It comprised 35 marks for MCQs, and 65 marks for questions requiring written answers. Our analysis uses the total test and examination marks.

5. RESULTS

This section presents the results of the statistical analyses performed for each of the five measures of activity. We have highlighted all p-values less than 0.05, indicating that we are at least 95% confident that the mean mark of the MPA students is greater than the mean mark of the LPA students.

5.1 Number of questions created (Q)

There is significance between the examination performance of the MPA and LPA students in three of the four quartiles (Table 4).

5.2 Number of questions answered (A)

There is significance between the examination performance of the MPA and LPA students in all four quartiles (Table 5).

Students in the MPA group of the fourth quartile show an increase of over 10 marks over the LPA students, the difference between a clear fail and a marginal pass.

5.3 Total length of comments (C)

There is significance between the examination performance of the MPA and LPA students in all four quartiles (Table 6).

Qt	Act	Avg Q	Test	Exam	Test p-value	Exam p-value
1	MPA	4.7	85.9	91.2	0.3364	**0.0478**
	LPA	1.5	85.4	88.8		
2	MPA	3.3	67.9	76.3	0.1318	0.1043
	LPA	1.2	66.8	73.7		
3	MPA	3.8	48.8	62.8	0.3859	**0.0458**
	LPA	1.3	48.5	58.4		
4	MPA	4.1	26.4	40.3	0.2966	**0.0164**
	LPA	1.0	25.5	33.4		

Table 4: Test and exam scores of MPA students compared with LPA students with respect to writing questions.

Qt	Act	Avg A	Test	Exam	Test p-value	Exam p-value
1	MPA	64.0	86.1	91.5	0.2248	**0.0146**
	LPA	13.2	85.3	88.5		
2	MPA	52.6	68.4	77.0	**0.0201**	**0.0270**
	LPA	8.9	66.3	73.0		
3	MPA	62.9	49.0	63.8	0.2759	**0.0072**
	LPA	9.0	48.4	57.4		
4	MPA	57.4	27.3	43.7	**0.0421**	**0.0000**
	LPA	6.9	24.6	30.1		

Table 5: Test and exam scores of MPA students compared with LPA students with respect to answering questions.

Qt	Act	Avg C	Test	Exam	Test p-value	Exam p-value
1	MPA	1968.3	86.4	91.9	0.0949	**0.0027**
	LPA	193.2	85.0	88.1		
2	MPA	1096.3	67.3	75.9	0.5211	0.1903
	LPA	25.6	67.4	74.1		
3	MPA	1263.7	48.6	62.0	0.5461	0.1479
	LPA	28.8	48.7	59.2		
4	MPA	1126.7	27.1	41.3	0.0685	**0.0028**
	LPA	4.4	24.8	32.4		

Table 6: Test and exam scores of MPA students compared with LPA students with respect to writing comments.

Students in the top quartile clearly write more detailed comments, on average, than those in the other quartiles. Even in the LPA groups, the average comment length of students in the top quartile is six times that of the next highest average.

5.4 Number of active days (D)

Qt	Act	Avg D	Test	Exam	Test p-value	Exam p-value
1	MPA	6.1	86.4	91.5	0.0921	**0.0146**
	LPA	2.0	85.0	88.5		
2	MPA	4.7	67.8	76.6	0.1798	0.0564
	LPA	1.3	66.9	73.4		
3	MPA	5.3	48.8	63.0	0.3729	**0.0377**
	LPA	1.4	48.5	58.2		
4	MPA	4.8	27.8	43.3	**0.0080**	**0.0000**
	LPA	0.8	24.1	30.5		

Table 7: Test and exam scores of MPA students compared with LPQ students with respect to the number of active days.

There is significance between the examination performance of the MPA and LPA students in three of the four quartiles, with the 2nd quartile difference approaching significance (Table 7).

5.5 Combined measure (CM)

Qt	Act	Avg CM	Test	Exam	Test p-value	Exam p-value
1	MPA	32.4	86.7	92.3	**0.0293**	**0.0005**
	LPA	18.5	84.7	87.7		
2	MPA	28.9	68.2	77.8	0.0559	**0.0030**
	LPA	13.6	66.5	72.2		
3	MPA	29.7	48.8	63.2	0.3728	**0.0256**
	LPA	13.8	48.5	58.0		
4	MPA	28.2	27.1	42.0	0.0716	**0.0006**
	LPA	10.9	24.8	31.7		

Table 8: Test and exam scores of MPA students compared with LPA students with respect to the combined metric.

There is significance between the examination performance of the MPA and LPA students in all four quartiles (Table 8).

In summary, there is strong evidence that students who are most active with PeerWise also perform better in final exams than their less active counterparts, for the activity measures and quartiles shown in Table 9.

	significance level	
	1%	5%
Q	4th	1st, 3rd and 4th
A	3rd and 4th	All quartiles
C	4th	1st and 4th
D	4th	1st, 3rd, and 4th
CM	1st, 2nd and 4th	All quartiles

Table 9: Summary of significant results by quartile

6. THE EFFECT OF PEERWISE ON EXAMINATION RESULTS

Our first analysis used the results of the mid-semester test to divide the students into quartiles (and thus implicitly into four pre-PeerWise use achievement groups). Our second analysis looked at the correlation between PeerWise activity and examination results (regardless of previous performance), and in particular, the results from multiple choice examination questions, and the results from the other examination questions requiring written answers. We used two measures of exam performance: the percentage mark for the multiple choice questions on the final examination (M) and the percentage mark for the questions on the final examination that were not multiple choice questions (but covered the same material) (E).

6.1 Hypotheses

We used a correlation coefficient test to see whether there was any relationship between the activity measures and the performance on the MCQ and non-MCQ examination questions.

- H0a: There is no relationship between the extent of a student's PeerWise activity and that student's performance on unseen multiple choice questions.
- H1a: There is a positive relationship between the extent of a student's PeerWise activity and that student's performance on unseen multiple choice questions, such that increased activity results in higher marks.
- H0b: There is no relationship between the extent of a student's PeerWise activity and that student's performance on unseen examination questions that are not multiple choice questions.
- H1b: There is a positive relationship between the extent of a student's PeerWise activity and that student's performance on unseen examination questions that are not multiple choice questions, such that increased activity results in higher marks.

Note: these research hypotheses are directional (i.e. permit one-tailed tests of significance)

6.2 Results

For each of the five activity measures, we performed a correlation analysis between the activity measure and the two examination performance measures. The results are shown in Table 10: the bold values show significance at the 5% level.

There is significant correlation between all five measures and the performance on the examination MCQs: this is not surprising, as PeerWise gives students' extensive practise in, and knowledge of, MCQs.

The correlation results for the written examination questions are more interesting, as there is no reason to expect that extensive PeerWise experience with MCQs would help in performance of written questions, unless such experience has led to a deeper understanding of the material. Only the measures of total length of comments, days active, and the combined measure are significantly related to performance in

	Performance in final MCQ examination (M)		Performance in final non-MCQ examination (E)	
	Correlation coefficient (r)	p-value	Correlation coefficient (r)	p-value
Number of questions created (Q)	0.105	**0.0121**	0.067	0.0757
Number of questions answered (A)	0.108	**0.0102**	0.046	0.1625
Total length of comments (C)	0.116	**0.0063**	0.106	**0.0115**
Number of active days (D)	0.189	**0.0000**	0.168	**0.0001**
Combined measure (CM)	0.321	**0.0000**	0.331	**0.0000**

Table 10: The correlation coefficients between five activity measures mid-semester final MCQ and non-MCQ performance, together with their probability values.

examination questions that required written answers. This suggests that it is not merely the activities of creating and answering MCQs that result in improved non-MCQ performance, but a high engagement with PeerWise (as evidenced by comments and activity days). This engagement thus suggests the development of a deeper level of understanding, enabling improved exam performance.

The combined measure gives a strong correlation for all examination performance, as shown in the scatter plots in Figure 1. While the correlation of the individual measures are statistically significant, the high correlation of the combined measure indicates that it is the *combination* of all four types of activity that is best related to exam performance.

7. DISCUSSION

These are encouraging results, showing that for most quartiles PeerWise activity is strongly related to exam performance, and that higher engagement in the use of PeerWise appears to foster deep learning (and hence higher non-MCQ examination marks).

These results raise further questions about students' engagement in the course, and the choice of the activity measures.

7.1 Were students who did not use PeerWise at all still engaged in the course?

The purpose of comparing the test marks of the MPA and LPA students was to verify that we were considering groups of students of equivalent abilities prior to their use of PeerWise. For some of our measures of activity there are significant differences between the mid-semester pre-PeerWise test marks of the students. As PeerWise was not introduced until after the test, this indicates that to a certain degree some of the activity measures separate the strong students from the weak students even within an achievement quartile.

A careful examination of the data shows that in each quartile, there are some students who have not participated with PeerWise at all according to the relevant metric. Table 11 summarises the number of students in each quartile who have no level of PeerWise participation according to each of the metrics, the "No PeerWise Activity" (NPA) students.

Clearly, all of these non-participating students will fall in the LPA group for each quartile.

An interesting question is whether these non-participating students were still active in other areas of the course. Certainly they all still sat the exam (non-sitting students were not included in our analysis). Other assessed activities in

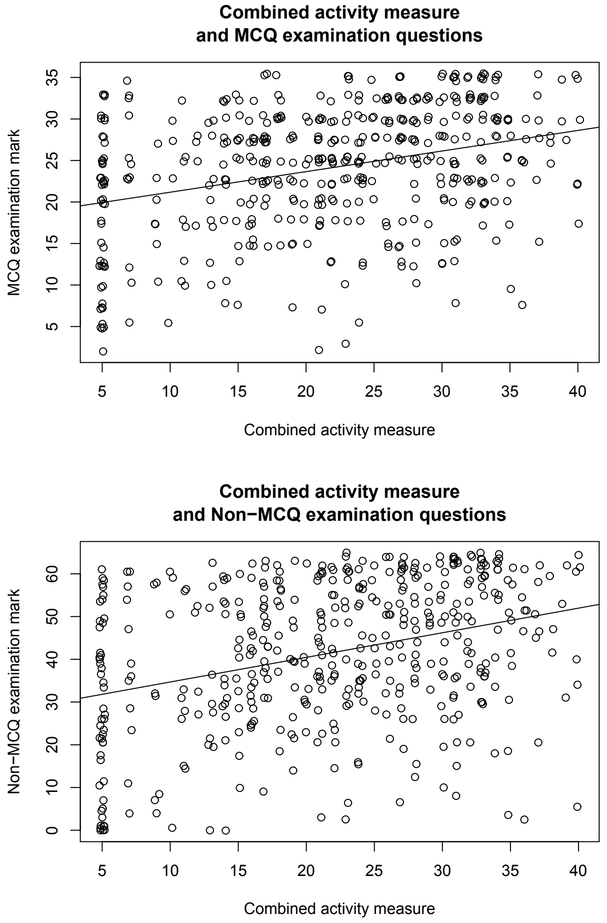

Figure 1: Scatter plots and trend lines for the combined activity measure and examination questions

Qt	Q	A	C	D
1	14	7	14	7
2	22	10	38	10
3	19	12	32	10
4	29	22	50	10

Table 11: Number of NPA students for each quartile, for each activity measure.

the course include ten supervised lab sessions throughout the semester and five unsupervised projects. Table 12 summarises the total number of students in each quartile who did not submit the projects, or attend the lab sessions during the period when PeerWise was being used.

Qt	Labs	Projects
1	1	2
2	0	6
3	3	7
4	10	22

Table 12: Number of students in each quartile who have not participated in labs or projects

This indicates that, particularly in the top 3 quartiles, the majority of students who did not participate in PeerWise were still actively engaged in the other assessed activities in the course. In the lowest quartile, 8% of students were no longer attending labs, and nearly 20% were not submitting their project work in the second half of the semester when PeerWise was used.

7.2 Were the metrics of activity appropriate?

Although we performed the analyses using individual metrics, the metrics themselves are not independent. For example intuitively, the greater the number of days of activity with PeerWise, the greater then number of answered questions is likely to be. Also, the number of comments contributed is bounded from above by the number of questions answered (as only one comment can be written per question answered).

Table 13 shows the correlation coefficients between all the activity measures, all of which are significant at the 1% level of significance.

These high correlations mean that attributing increased performance within quartiles between the mid-semester test and the final examination to only one measure of activity may be inappropriate, as taking a high measure of one activity is implicitly also taking a high measure of another activity.

However, when considering students' achievement in the section of the final examination that was not multiple choice questions, it is clear that separating the measures yields interesting results, regardless of the correlation between them. It is interesting to note that it is the number of active days (rather than the number of questions created) and the length of comments written (rather than the number of questions answered) that had an effect on students' performance on the non-MCQ exam questions.

7.3 Why was there less effect on the second quartile?

	A	C	D
Q	0.416	0.322	0.526
A		0.504	0.702
C			0.426

Table 13: Correlations between the different activity measures, all of which are significant at 1%.

Students in the second quartile did not show significant differences in examination results with respect to writing questions, writing comments (along with the third quartile), or days active. In contrast, the top and bottom quartiles showed significant activity related improvements over all the measures we considered.

We can only surmise as to why this phenomenon arose. It is apparent from browsing the PeerWise database that top students used the system effectively, and engaged in deep, reflective learning. Weak students had the most to gain from a drill-and-practice interaction with the predominantly basic course material accumulated in PeerWise. However, students of mid-ability, who had presumably already mastered the basics, were not at a level that allowed them to engage more deeply with PeerWise.

If this analysis is correct, the challenge will be to develop a tool that scaffolds learners at all ability levels.

8. CONCLUSIONS

Active use of PeerWise is strongly related to students' grades in both the multi-choice and written sections of the final examination. The improvements in the written sections imply that PeerWise use may have resulted in deep learning, rather than simply coaching students into better MCQ technique. Our analysis of different measures of PeerWise activity suggest that, in addition to time-on-task, voluntary engagement through the question discussion forum is a strong contributor to this improvement.

The benefits of PeerWise are not confined to students of just high or just low ability. We see improvements across all performance quartiles, and most consistently in the top and bottom groups. There is some evidence to suggest that PeerWise is of less benefit to mid-ability students, which raises the challenge of developing variations on this tool capable of scaffolding learners at all ability levels.

We plan to replicate this experiment in the first semester 2008 under similar conditions, and are also looking to collect similar data from courses in other disciplines and institutions that adopt PeerWise.

9. REFERENCES

[1] N. Arthur. Using student-generated assessment items to enhance teamwork, feedback and the learning process. *Synergy*, 24:21–23, 2006. www.itl.usyd.edu.au/synergy.

[2] M. Barak and S. Rafaeli. On-line question-posing and peer-assessment as means for web-based knowledge sharing in learning. *International Journal of Human-Computer Studies*, 61:84–103, 2004.

[3] M. Birenbaum and R. A. Feldman. Relationships between learning patterns and attitudes towards two assessment formats. *Educational Research*, 40(1):90–97, 1998.

[4] D. Boud and N. Falchikov. Quantitative studies of student self-assessment in higher education: a critical analysis of findings. *Higher Education*, 18(5):529–549, 1989.

[5] P. Brusilovsky and S. Sosnovsky. Engaging students to work with self-assessment questions: a study of two approaches. In *ITiCSE'05: Proceedings of the 10th Annual SIGCSE Conference on Innovation and*

Technology in Computer Science Education, pages 251–255, New York, NY, USA, 2005. ACM.

[6] B. Collis and J. Moonen. *Flexible Learning in a Digital World: Experiences and Expectations*. Kogan Page, London, 2001.

[7] P. Denny, A. Luxton-Reilly, and J. Hamer. The PeerWise system of student contributed assessment questions. In Simon and M. Hamilton, editors, *Tenth Australasian Computing Education Conference (ACE 2008)*, volume 78 of *CRPIT*, pages 69–74, Wollongong, NSW, Australia, 2008. ACS.

[8] P. Denny, A. Luxton-Reilly, and J. Hamer. Student use of the PeerWise system. In *ITICSE'08: Proceedings of the 13th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education*, pages 73–77, New York, NY, USA, 2008. ACM.

[9] S. Chang. et al. AGQ: a model of student question generation supported by one-on-one educational computing. In *CSCL'05: Computer Support for Collaborative Learning*, pages 28–32. International Society of the Learning Sciences, 2005.

[10] N. Falchikov and D. Boud. Student self-assessment in higher education: a meta-analysis. *Review of Educational Research*, 59(4):395–430, 1989.

[11] M. Fellenz. Using assessment to support higher level learning: the multiple choice item development assignment. *Assessment and Evaluation in Higher Education*, 29(6):703–719, 2004.

[12] S. A. Horgen. Pedagogical use of multiple choice tests — students create their own tests. In P. Kefalas, A. Sotiriadou, G. Davies, and A. McGettrick, editors, *Proceedings of the Informatics Education Europe II Conference*. SEERC, 2007.

[13] C. Lutteroth and A. Luxton-Reilly. Flexible learning in CS2: a case study. In *(to appear) Proceedings of the 21st Annual Conference of the National Advisory Committee on Computing Qualifications*, Auckland, New Zealand, 2008.

[14] D. Nicol. E-assessment by design: using multiple-choice tests to good effect. *Journal of Further and Higher Education*, 31(1):53–64, 2007.

[15] D. Nicol and D. Macfarlane-Dick. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2):199–218, 2006.

[16] K. Scouller. The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education*, 35(4):453–472, 1998.

[17] K. Topping. Peer assessment between students in colleges and universities. *Review of Educational Research*, 68(3):249–276, 1998.

[18] E. Wenger. *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press, 1998.

[19] J. Whalley, R. Lister, E. Thompson, T. Clear, P. Robbins, P. Kumar, and C. Prasad. An Australasian study of reading and comprehension skills in novice programmers, using the Bloom and SOLO taxonomies. In *ACE'06: Proceedings of the 8th Australasian Conference on Computing education*, pages 243–252, Darlinghurst, Australia, 2006. Australian Computer Society, Inc.

[20] F. Yu, Y. Liu, and T. Chan. A web-based learning system for question posing and peer assessment. *Innovations in Education and Teaching International*, 42(4):337–348, 2005.