Engaging Students and Evaluating Learning Progress using Collaborative Exams in Introductory Courses

Timothy T. Eaton[1]

ABSTRACT

Collaborative exams, with subsections in which students have an opportunity to work with each other, are not yet widely used in introductory geoscience courses. This type of exam structure, with other participatory teaching strategies, was employed in two different courses, and results were found to provide a sensitive and revealing tool for analyzing the progress of students' individual and collaborative learning throughout the semester. A somewhat different implementation of the collaborative exams in each course showed that overall student performance was significantly improved compared to performance in the individual part, even for middle and highest-achieving thirds of the student population, and progressive improvements in performance were followed through the semester. The implementation of collaborative exams in the first course involved an aspect of exam grading that provided an incentive for collaboration: the "jackpot effect", which provided insight into the dynamics of peer interaction. The simpler implementation in the second course used a different approach in which the collaborative tests were less important to the total class grade, but also showed improvement in students' individual performance. Wider application of these methods could make a critical difference in reversing student apathy toward science in colleges and universities.

INTRODUCTION

"The instructor is a good professor, just the course itself was not of interest to me." This sentiment, expressed on an end-of-semester teaching evaluation for an introductory environmental science course at Queens College, unfortunately characterizes a large fraction of student attitudes towards science at many U.S. colleges and universities. The National Research Council has recognized the critical need for science literacy among the United States citizenry in the 21st century (National Research Council, 1996). More recent studies have highlighted the connection between science literacy and national security or the economic future of this country (National Academy of Science, 2005) and proposed a national action plan for the U.S. education system to meet these needs (National Science Board, 2007). The U.S. system of primary and secondary education does not do a particularly good job of preparing students for learning science at the college level, as indicated by recent scores on the international TIMSS tests compared to other countries (Gonzales et al., 2004), yet undergraduate college is the last stage in formal education for most Americans. It is imperative that introductory courses in science at the college level generate enough interest to captivate what has been called the "second tier" (Felder, 1993) of students who have the ability to do so, but elect not to enter careers in fields of science. Moreover, even if students are not interested in science as a profession, a scientifically educated citizenry is essential to future decisionmaking through our representative democratic system of government. Wider application of innovative college teaching methods is needed to combat the current apathy about science expressed by the student quoted above. This paper focuses on innovative collaborative exam implementation and results for two geoscience classes at Queens College.

The college science requirements at Queens College benefit the School of Earth and Environmental Sciences at Queens College because students perceive, somewhat incorrectly, that "environmental science" or "geology" must be easier than the "hard" traditional sciences of Physics, Chemistry and Biology, so they flock to our introductory courses that meet the requirement of one lab course and one non-lab course in science. Two of these geoscience courses will be discussed here. ENSCI111: Introduction to the Environment is a lecture and laboratory course that fills the dual need for an introduction to the major courses, and satisfies the college requirements for a science lab course. GEOL25: Natural Resources and the Environment is a lecture-based service course that fills the requirement for a non-lab science course, and attracts mostly non-majors. These courses have been growing in enrollment over the last few years, with ENSCI111 reaching up to 400-500 students divided into three lectures and many lab sections. GEOL25 is a more modestly sized class, and has typical enrollments from 50 to over 100 students a semester. In fall 2007, the author taught an ENSCI111 lecture section, and in fall 2008, he taught GEOL25, and results are presented from both courses.

The challenge for college science instructors is to present to students three basic elements: 1) some understanding of the basic tools scientists use to investigate the world around us; 2) some understanding of the physical, chemical and biological processes that operate in the universe; and 3) some way of connecting the first two elements to students' everyday lives. The first two elements involve such skills as applying critical analysis to the wealth of information that is now available to the public through the media and internet, being able to interpret quantitative data presented in graphs and tables, and having some understanding of application of the scientific method and peer review, by which scientists advance the state-of-knowledge of scientific inquiry. As for the third element, students become more invested in course material if they are encouraged to discuss it among

---

[1] School for Earth and Environmental Science, Queens College CUNY, 65-30 Kissena Blvd, Flushing, NY 11367; Timothy.Eaton@qc.cuny.edu

themselves. Addressing the challenge posed by these considerations was a primary motivation in employing collaborative exams and other elements of class participation in the two courses. These innovations, mainly collaborative or pyramid exams (Cortright et al., 2003), but also exercises known as ConcepTests (McConnell et al., 2006), and learning cycle/think-pair-share (Reynolds and Peacock, 1998) have been described individually elsewhere, but their implementation here is combined in a way not presented before. While commonly discussed in professional development workshops like those sponsored by the National Science Foundation (http://serc.carleton.edu/NAGTWorkshops/index.html), these techniques have been described in a few publications (McConnell et al., 2003; Schwab, 2005; Yuretich et al., 2001), but are only slowly entering the mainstream of college education for large classes, particularly in the geosciences.

In many cases, while collaborative exam-taking has been demonstrated to improve student learning in small, carefully controlled groups (Rao et al., 2002; Zipp, 2007), as well as in specialized settings (Shindler, 2004) and in an online framework (Shen et al., 2008), it has not yet been described in sufficient detail for many faculty to begin adopting as standard practice. In particular, the "collaborative" effect on grades for students at different skill levels in the course, and the ramifications of collaborative work on achievement in successive exams within a course have not been investigated, to the author's knowledge, in any detail. Furthermore, specific effects on grades of an exam correction method presented by Yuretich et al. (2001) have significant advantages as an incentive for the inquiry and student interaction associated with collaborative activities. Additional findings about various methods of encouraging class participation have proved to be useful in maintaining attendance, stimulating student interest, and gauging student comprehension of course material. The combination of all these methods and their implementation in practice in a large class setting has provided a perspective that is likely to be helpful to instructors contemplating their adoption in their own courses.

METHODS
Collaborative exams

The current structure of the two geoscience courses at Queens College led to different implementations of the collaborative exam method. The lecture part of the ENSCI111 course has traditionally been divided into five parts, focusing roughly on the topics of Science/Critical Thinking and the Environment, Climate Change/Air Quality, Water Resources and Quality, Ecosystems and Human Applications, and Energy Resources/Waste Disposal. Different emphases within these areas depend on the professors teaching the course, who have had primary research expertise in climatology, oceanography, soil microbiology and hydrology in recent years. Student evaluation has consisted of five multiple-choice exams corresponding to each of the five parts, written homework assignments focusing on reading comprehension of issue topics, class participation and occasionally extra-credit writing assignments. On the other hand, the GEOL25 course has had more of a focus on natural resources, minerals and energy. It has traditionally had only a midterm and a final exam (both multiple-choice), and been taught at a lower level than the ENSCI111 major course. This provided the opportunity to make the collaborative method much less high-stakes by using it only for intermediate, shorter multiple-choice tests (like quizzes) called Comprehension Evaluations, and use the non-collaborative midterm and final as a way of evaluating the effect.

Multiple choice exams have often been derided as akin to a guessing game or worse yet, reducing important and complex information to trivial details. Faculty who have taught large lecture classes (>100 students) without the benefit of teaching assistants know that because of machine-grading, multiple-choice exams are often one of the only viable options for student evaluation. The challenge then becomes how to use this tool in the most effective way to test how students have learned. To adapt standard multiple choice tests for the two courses here, three strategies were used: 1) reducing the number of questions so they could be made more complex, and students would have more time to spend on each; 2) incorporating image or graphical information on some questions as a basis for students to select different responses; 3) adding a collaborative section of the test on which students can work together. In this way, the level of student evaluation was raised from the basic knowledge category of Bloom's taxonomy (Bloom, 1956) to higher levels which are characterized by comprehension or interpretation, application to novel situations, and analysis or logical reasoning. The visual aspect of the images and diagrams on some test questions also addresses one of the neglected poles of the key visual-verbal dimension of the Index of Learning Styles (ILS) (Felder and Spurlin, 2005), that many students may have a preference for. The practical implementation of these strategies is described next.

In the ENSCI111 course, four of the five collaborative exams (the lowest grade was dropped) accounted for a large fraction (70%) of the course grade, whereas for the GEOL25 course, five of the six (the lowest grade was dropped) comprehension evaluations (CE's) accounted for only 30% of the course grade. In both cases, the collaborative tests consisted of 2/3 individual questions and 1/3 questions on which students had the option of collaborating, and were encouraged to work with the 4-5 students sitting nearby. This arrangement was more formalized in the GEOL25 class when students were assigned (alphabetically) to groups in which they were expected to work during the semester. Two separate answer sheets were used by each student, with the individual question answers turned in after about 45-50 minutes of the 75 minute period. The remaining collaborative questions were answered in the remaining time, and for the ENSCI111 course, consisted of repeated questions from the individual part of the exam, and for the GEOL25 class, different questions. Since the GEOL25 class was taught at a lower level, similar questions from

the collaborative tests were also used on the midterm and final exams, although numerical values were often changed and other minor modifications made. To minimize the temptation to cheat in a crowded lecture hall, three versions of the tests were used, with differently sorted individual questions on the first part, but the collaborative questions were listed in the same order on the second part. Students were provided with a sample of the cover page with instructions, and a full explanation for the exam format beforehand. After each test, students were provided with answer keys on electronic reserve and complete individual results, including how many points were assigned for each part and why. Explanatory comments were also provided for many of the correct answers indicated.

An important component of the collaborative tests was how the questions were graded, based on ideas proposed by Yuretich et al. (2001). The simpler case is for the GEOL25 course, where questions were all graded in the same way, and mean percent correct answers were tabulated for each part of the tests for different cohorts of students. The results provided a basis for comparison to evaluate growth in individual student achievement on the non-collaborative midterm and final exams. For the earlier ENSCI111 course, a more complex system was used to encourage peer interaction. Students are understandably concerned about collaboration at first, so the instructor specified that they should feel free to change their answer for the repeat questions on the second part *if* they are convinced that a different answer is best (there was only one best answer). Students were not penalized if they initially had the correct answer (on the individual part), but were misled and put a wrong answer for the same question on the collaborative part. This was done by weighting each question (total for both parts) equally, but crediting points even for wrong answers on the second part *provided they had already answered those same questions correctly* on the first part. The four possible cases for answers to duplicated questions (Table 1) include an interesting mechanism of eliciting individual performance (referred to as the "jackpot effect") that provided a substantial incentive to peer learning and subsequent individual achievement. In both courses, the comparison of achievement on both individual and collaborative parts

TABLE 1. POINT ASSIGNMENT FOR DUPLICATED QUESTIONS ON ENSCI111 COLLABORATIVE

|	Part 1 question	Part 2 question	Total for exam
Case 1	Correct = 1 point	Correct = 1 point	2 points
Case 2	Wrong = 0 points	Wrong = 0 points	0 points
Case 3	Wrong = 0 points	Correct = 1 point	1 point
Case 4 (jackpot effect)	Correct = 1 point	Wrong = 0 point	1+1= 2 points[1]

Notes:
[1]Point credited for part 2 as well as part 1 since initial answer was correct

of the tests, as well as longitudinal analysis of mean performance during the semester provides insight into the benefit of collaborative work. Mean percent results were calculated for each of the exam parts and these data were split out according to the upper third, middle third and lower third of the grades for each test. For the ENSCI111 class, the statistical significance of the difference in mean grades by student between the first (individual) part and the combined (individual and collaborative) parts for each cohort in each test was evaluated using a student's paired t-test. For the GEOL25 class, the statistical significance was similarly evaluated of the difference in mean grades for the student population between the comprehension evaluations and the midterm and final exam. Other aspects of test performance were also quantified, such as inter-exam grade improvement and what proportion of students in each cohort benefited from the collaborative exam format.

Lectures and other class participation

One of the challenges in interdisciplinary introductory courses such as ENSCI111 and GEOL25 is reconciling the amount of material presented with the need to explain concepts in detail. The lectures were presented using digital slide (Powerpoint) presentations, but the number of slides in a lecture (~20-30) was reduced compared to standard practice in previous semesters, to allow more time to discuss the concepts presented. Digital slide presentations in the geosciences have been effectively used to promote active learning because they allow presentation of imagery and promote skills of visual observation (Reynolds and Peacock, 1998). In this class, presentations consisted largely of photographic images, diagrams, charts and graphs, supplemented with a minimal amount of text summarizing basic points. This allowed the instructor to present data and discuss figures and graphs that students also had access to in their class readings and on electronic reserve.

Slide imagery has important benefits over the traditional chalk-on-blackboard technique for lecture presentation in earth sciences. This lecture presentation method, based more on images than words, builds on the exceptional reliance humans have on their visual perception for absorbing information. It also enables the instructor to model how scientists analyze and interpret data about the world around us. Digital presentation can also be very effective in illustrating changes over time using sequential imagery – a technique that is particularly well suited to topics of global climate change. For example, the reduction of the annual coverage of summer sea ice in the Arctic Ocean since 1979, or the recent collapse of the West Antarctic Ice Sheet is made particularly compelling in this way. Compilations of flash animations are increasingly available online to help explain complex topics such as how combined sewer systems in large urban areas like New York City contribute to poor water quality in the surrounding harbor. Such images, diagrams or graphs were incorporated into some questions on the collaborative tests in each course (as well as the midterm and final in GEOL25), and used to evaluate student retention and
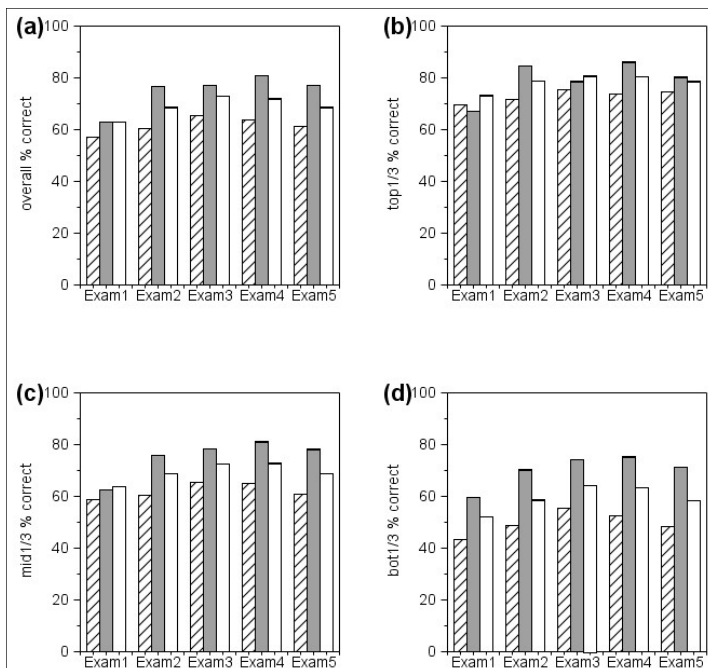
FIGURE 1. Mean percent exam grades in ENSCI111 by exam achievement cohort and by sections of the exam. In each set of three bars, the first (hachured) bar represents the grade on the first (individual) part of the exam, the middle bar (gray) represents the grade on the second (collaborative) part of the exam, and the third bar (white) represents the combined exam grade, including the "jackpot effect" points credited (see text). a) entire class, b) highest-achieving 1/3 of students, c) middle-achieving 1/3 of students, and d) lowest-achieving 1/3 of students.

comprehension of course material.

Lack of attendance to lecture is a common problem in large enrollment classes (Moore, 2003), and attendance for the two classes had been poor in previous semesters, although it was not systematically recorded. To improve this aspect of the courses, the instructor made it clear that class participation would be part of the course evaluation, accounting for 10% of the final grade. Several methods were used to encourage participation. The more limited number of slides per lecture allowed repeatedly stopping in the middle of lecture and asking students about specific personal experience with various environmental/ geoscience issues or topics covered that day. Due to the limited ability to interact with individuals in such large classes, other techniques involving specific point assignments for in-class written exercises, were also introduced to monitor attendance as well as evaluate participation.

In both classes, on at least six different but unannounced occasions scattered throughout the semester, students were asked to hand in a small assignment worth from 5-10 points. No points were awarded if students failed to hand in the activity assignments. These ranged from simple responses on a 5x8" card to full-blown think-pair-share activities (Macdonald and Korinek, 1995). Several occasions (at the beginning of class to encourage promptness) were used to

assess student comprehension of the previous lecture by asking them to list one or two topics which they understood well, and one or two topics which they didn't understand (the "muddiest point" technique). This feedback technique was also used after the first ENSCI111 exam to evaluate student attitudes about the collaborative format. Other opportunities in both ENSCI111 and GEOL25 were used to elicit student reflections regarding class topics. For example, when modern agriculture was discussed in ENSCI111, students had been asked beforehand to identify several perishable goods from their refrigerator and how far they had been transported from their location of production. In GEOL25, exercises included identifying caloric content of foods belonging to the major food groups (grains, pulses, meats, tubers) that are eaten around the world, and participating in hypothetical oil exploration and discovery. Another activity in both courses focused on fuel efficiency of different vehicles driven by students, presented as a histogram in discussion about energy resources and conservation.

RESULTS

Results from these innovations in teaching these introductory courses are semi-quantitative and quantitative. Semi-quantitative observations in ENSCI111 include a level of attendance ranging between 50% and over 90%, but averaging around 73% based on turn-in ratios of class participation and homework assignments (during class for the most part) as well as exam presence. Since detailed records of attendance are not kept from previous offerings of this course, it is difficult to quantify improvement here, however anecdotal evidence suggests
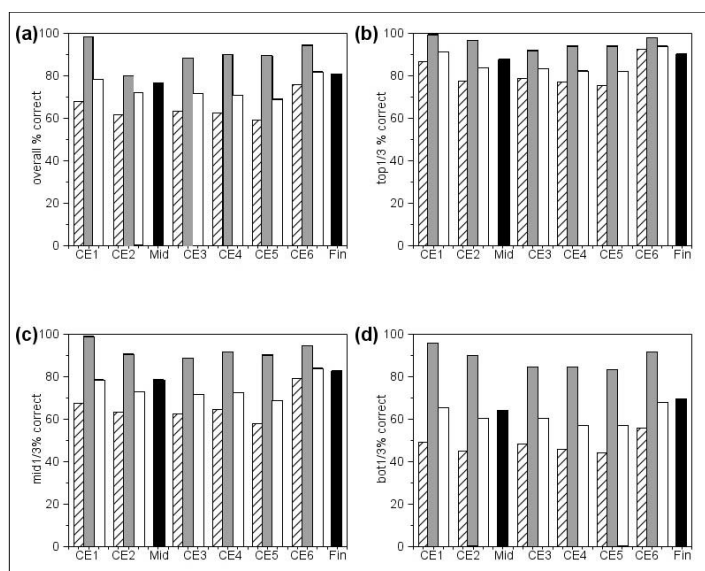
FIGURE 2. Mean percent test grades in GEOL25 by test achievement cohort and by sections of the test. Legend and figure parts are the same as in Figure 1 except that black bars indicate the mean grades on the midterm and final exams, which were not collaborative exams. Only two comprehension evaluation (CE) tests could be given before the midterm because of the extensive fall break during the semester.

TABLE 2. WITHIN-EXAM IMPROVEMENT IN ENSCI111: PERCENT CORRECT DIFFERENCE BETWEEN TOTAL EXAM GRADE AND PART 1 GRADE

Cohort	Exam 1	Exam 2	Exam 3	Exam 4	Exam 5
Entire	5.75% (n=112)	8.25% (n=115)	6.89% (n=109)	8.49% (n=102)	7.40% (n=107)
Top 1/3 of class	3.77% (n=37)	6.71% (n=38)	5.26% (n=36)	6.80% (n=34)	4.20% (n=36)
Mid 1/3 of class	5.10% (n=38)	8.16% (n=39)	6.65% (n=37)	7.91% (n=34)	7.99% (n=35)
Bottom 1/3 of class	8.39% (n=37)	9.89% (n=38)	8.75% (n=36)	10.76% (n=34)	10.01% (n=36)

that attendance has been increased somewhat, despite the constraints of the assigned lecture hall, which barely seats 115 students. Similar results were noted for the GEOL25 class. Compared to previously reported rates of less than 70% from other large introductory lecture courses elsewhere (Moore, 2003; Yuretich et al., 2001), an attendance rate averaging 73% seems to represent progress.

Students initially had an almost uniformly enthusiastic response in the feedback class activity to the ENSCI111 collaborative exam format, which was tempered somewhat as the semester progressed. Responses to the collaborative testing in GEOL25 remained positive. End-of-semester faculty evaluations by students consist of questions on a scale of 1 to 5 and mean scores suggest that on balance, students found reading and other assignments to be valuable, and enjoyed and learned a great deal in the course. However, they did find it moderately to very difficult, a little fast-paced, and a moderate to heavy workload. There were also quite a few complaints indicating that the students didn't fully understand exactly how the collaborative aspect in ENSCI111 worked, which prompted the simplification in the later GEOL25 class. However, numerous comments expressed enthusiasm for the classroom activities.

Collaborative exams

The slightly different implementations of the collaborative testing in the two courses provide insight into the effect of the peer learning that is accomplished. There are some overall trends that stand out for both courses. The major quantitative results from this study are presented in Figures 1 and 2, showing the mean percent test grades for each part of each test and the combined grade for each test. In the case of GEOL25, collaborative testing was not used for the midterm and the final. Obviously the grades were in virtually all cases better for the collaborative parts compared to the individual parts. There is a noticeable and highly significant (P*=0.001) increase in grade percentage in all cases between the individual part and the combined grade for each ENSCI111 exam, and between the mean grades on

sequential exams (Tables 2,3).

For the GEOL25 class, there are notable increases in mean grades between the individual and combined parts of the comprehension evaluations, and significant increases in mean percent grades between the individual parts of the CE tests and the midterm or final, although some improvements are at a lower level of significance (Tables 4, 5). With the exception of CE1 and CE6, the mean individual midterm and final exam grades are generally better than the mean grades for the combined parts of the preceding comprehension evaluations (Figure 2).

For both courses (Figures 1, 2), the general pattern is that the combined mean test scores are increased compared to the part 1 (individual) grades, but not up to the level of the mean scores on part 2 (collaborative). This pattern is best illustrated by the lowest achieving third of the ENSCI111 population grades on all exams (Figure 1d). For the ENSCI111 class, there are some interesting exceptions to this pattern where combined mean grades sometimes exceed the mean grades for the individual parts *and* the collaborative parts of the exam, but this never happens for the GEOL25 class (Figure 2).

Improvement during course

Many evaluations of innovative teaching methods compare student performance in previous course offerings to that for the semester in which they are implemented, but such an approach was not possible here. However, the use of the collaborative test format allows more refined monitoring of student progress during the semester itself, because it provides a mechanism for students to improve their learning from each other. One of the interesting results of this study is that the boost from the collaborative effect is present at all levels of student achievement, as shown in Figures 1 and 2. Another aspect of this progression is how the exam grades improve from exam to exam. To evaluate student progress in ENSCI111, the percentages of students in each cohort (high, mid and low-achieving 1/3 of exam-takers) who have significant improvement (defined as an increase in 5 percentage points, or at least one third of a letter grade: i.e. B- to B) in their grades from one exam to the next (Table 3) was calculated. In ENSCI111, students could

TABLE 3. SIGNIFICANT INTER-EXAM GRADE IMPROVEMENT ENSCI111 (+5 POINTS)

Cohort	*Exam 1-> next exam*	*Exam 2 -> next exam*	*Exam 3-> next exam*	Exam 4-> last exam
Top 1/3 of class	32.4% (n=37)	7.9% (n=38)	16.7% (n=36)	2.9% (n=34)
Mid 1/3 of class	50.0% (n=38)	38.5% (n=39)	37.8% (n=37)	23.5% (n=34)
Bottom 1/3 of class	75.7% (n=37)	68.4% (n=38)	52.8% (n=36)	38.2% (n=34)

TABLE 4. WITHIN-TEST IMPROVEMENT GEOL25: PERCENT CORRECT DIFFERENCE BETWEEN TOTAL TEST GRADE AND PART 1 GRADE

Cohort	CE 1	CE 2	CE 3	CE 4	CE 5	CE 6
Entire	10.61%(n=77)	10.26% (n=77)	8.38% (n=80)	8.16% (n=76)	10.02% (n=78)	6.23% (n=76)
Top 1/3 of class	4.57% (n=26)	6.47% (n=26)	4.44% (n=27)	5.47% (n=25)	6.22% (n=26)	1.73% (n=25)
Mid 1/3 of class	11.02% (n=25)	9.27% (n=25)	8.53% (n=26)	8.01% (n=26)	10.90% (n=26)	5.06% (n=26)
Bottom 1/3 of class	16.27% (n=26)	15.00% (n=26)	12.16% (n=27)	11.00% (n=25)	12.95% (n=26)	11.93% (n=25)

TABLE 5. PERCENT GRADE IMPROVEMENT FROM CE TESTS TO MIDTERM AND FINAL IN GEOL25 EXAM GRADE AND PART 1 GRADE

Cohort	CE1 -> Midterm	CE2 -> Midterm	CE3 -> Final	CE4 -> Final	CE5 -> Final	CE6 -> Final
Top 1/3 of class	14.3% (n=22)	21.0% (n=22)	20.1% (n=27)	18.0% (n=27)	27.2% (n=27)	4.9% (n=27)[1]
Mid 1/3 of class	11.8% (n=22)[1]	18.5% (n=22)	23.1% (n=26)	23.5% (n=26)	17.9% (n=26)	8.5% (n=26)[1]
Bottom 1/3 of class	5.5% (n=22)[1,2]	12.3% (n=22)	8.6% (n=27)[1]	13.6% (n=27)	18.6% (n=27)	2.7% (n=27)[3]

Notes:
[1] not significant to 0.001
[2] not significant to 0.01
[3] not significant to 0.05

drop one exam grade, and many elected to skip an exam, so the data compare the scores from one exam to the next that students took. For GEOL25, since the comprehension evaluations counted for much less of the total grade, the difference between the individual midterm and final exam mean scores by student, and the preceding individual scores on the comprehension evaluations was calculated instead (Table 5). The mean improvement by student between the first two comprehension evaluations and the midterm, and between the last four comprehension evaluations and the final was generally in the double-digit percentages except for the lowest achieving cohort, as well as specifically for CE6 compared to the final.

DISCUSSION

Any inferences of student learning based on multiple-choice tests must first assume that students are not merely guessing at all or most of the answers. If this were the case, then multiple-choice tests would be useless as an evaluation tool for academic learning. While a certain amount of guessing is inevitable, so the tool is imperfect, this type of test assumes that the percentage of correct answers is largely correlated with the amount of learning done by each student. In this context, the comparison between the results for the somewhat different implementation of the collaborative test approach provides an interesting perspective on the dynamics of student interaction as they learn the best ways of analyzing the test problems from their peers. One might expect the collaborative test format to benefit only the lowest-achieving students, because they would have access to the correct answers from their higher-achieving classmates during the collaborative part. However, in the case of tests in both courses, it is striking how at all three levels: lowest, middle and high-achieving students, the

mean grades increase in the combined (individual plus collaborative) parts compared to the individual part alone.

One might infer that the use of a collaborative section on a multiple-choice test serves only to inflate student grades, a common problem in college settings. One benefit of the collaborative testing scheme is that the author's experience suggests there doesn't appear to be any need to apply a "curve" or make any adjustment to the point scores to ensure that the class mean is near the expected 70% or C average. A corollary benefit is that retention of class material can therefore be tested at a more complex level on the scale of Bloom's taxonomy (Bloom, 1956) without having an inordinately high test failure rate. It is certainly possible that students simply put down the answer to questions on the collaborative sections on the basis of consensus among their peers, without really internalizing or learning the relevant information. However, the two ways of implementing the collaborative test technique in each course provide some evidence that the method does improve student learning over conventional testing.

In the ENSCI111 course exams, there was an incentive built in to the grading structure for the students to learn collaboratively, yet not be penalized if they knew the correct answer and were later misled by their peers. This incentive, referred to as the "jackpot effect", whereby students could obtain credit for questions for which they had earlier shown they knew the correct answer (Table 1), entails a more complex grading effort and duplication of questions on the individual and collaborative parts of the test. The relative percentages of students by cohort who obtained extra points on the collaborative part of the exams throughout the semester is presented in Figure 3. The effect of this is to encourage students to collaborate "intelligently" rather than just putting down the

consensus answer, and spurs them to learn how to best analyze and respond successfully to exam questions without being misled. A limit on the number of extra points (5) that could be so credited was instituted because a few students were deliberately choosing different answers simply to increase their odds of guessing correctly. Higher percentages indicate a greater willingness to change the answers for the same question from part 1 (individual) to part 2 (collaborative), and a lack of confidence in students' own answers. Variation in the percentages in Figure 3 may also be a response to the perceived difficulty of any exam, or a result of students becoming more accustomed to the exam format or more adept at using collaboration to improve their test performance or individual learning. The latter explanation is more probable because the general trend of increase in scores over the five exams during the semester (Figure 1) suggests that the exams were largely similar in difficulty (with the possible exception of Exam 5), so students were able to improve their collective (Figure 1) and individual (Table 3) performance over time.

Despite its complexity, this incentive ("jackpot") grading mechanism for the exams in ENSCI111 suggests that the collaborative structure enhances students' ability to improve their overall grade on the tests, rather than simply inflating grades. Consider that the mean number of exam questions by student for which points were added to compensate for the "right then wrong" pattern
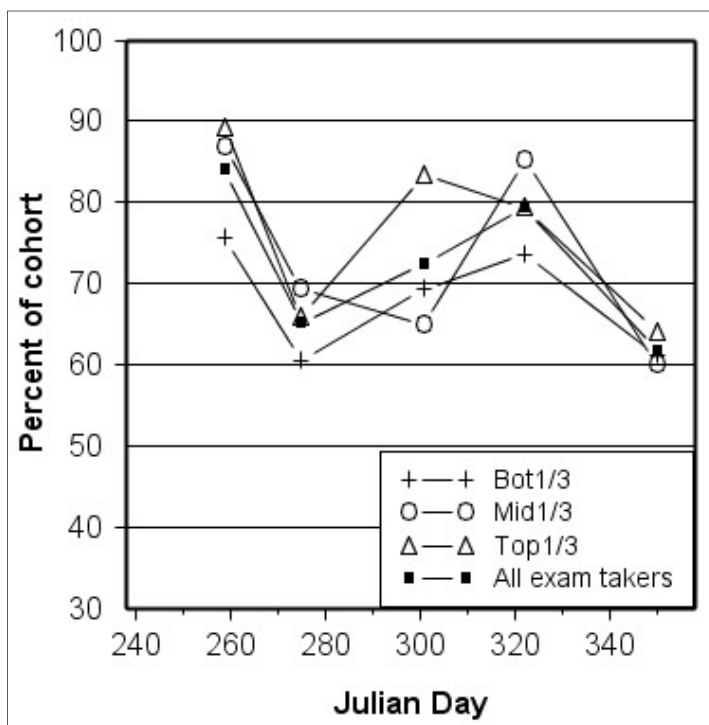

FIGURE 3. Variation in the proportion of beneficiaries of the "jackpot effect" (see text) in the ENSCI111 collaborative exam grading scheme as a percentage of different class cohorts. The top one-third achieving students, the middle-achieving students and the lower one-third achieving students (see Tables 2,3 for numbers) are identified relative to all students taking each exam.

was slightly higher than 2 for the first exam, then declined over the different exams to 1.2 for Exam 5. A conservative (high) estimate of the mean percentage of points gained is therefore 100*2/58 (exams had 58 questions each except for one with 59) or less than 3.5%. It is clear from Figure 1 and Table 2 that the actual improvement in exam scores from the individual to the combined grades is in all cases greater than this. Of course, this calculation based on mean data generalizes the details for each student, and some students undoubtedly benefited even more.

The alternate approach to collaborative testing for GEOL25 (Figure 2) simplifies the grading burden and omits the "jackpot" incentive for collaboration. However, the data for the non-collaborative midterm and final also show that students have improved their achievement ability (Table 5) compared to the individual parts of the preceding collaborative comprehension evaluations (CE tests). By weighting the bulk of the course grade on the non-collaborative midterm and final exams, it is clearer that student achievement is based on their own ability rather than collaboration. The collaborative evaluation (CE) tests therefore provide a lower-stakes way for students to learn from each other than in the earlier approach used for the ENSCI111 course. The sheer number of tests (CE's as well as midterm and final) seemed to be a little too much in GEOL25, so the number of CE tests has since been reduced to five in the current semester.

CONCLUSIONS

Two different implementations of the collaborative (or pyramid) test method were used in introductory geoscience courses: ENSCI111 and GEOL25 respectively. Results of a comparative analysis of student achievement on individual and collaborative sections, as well as a longitudinal analysis of progressive performance during the semester, indicate that the collaborative testing technique improves student learning. The benefits of the combination of the collaborative testing and other innovative teaching methods that enhanced student participation go far beyond the sum of the parts (each individual test or class exercise).

The clearest evidence for the benefits of collaborative learning come from analysis of the mean performance by student in both individual and collaborative test parts, split out by upper, middle and lower-achieving cohorts of the class, and the results in successive tests during the semester. While the lowest-achieving cohort benefited most from the collaboration, the middle and upper cohorts also experienced significant improvements in their grades compared to their individual performance. In addition, there is a notable increase in student performance from one collaborative exam to the next in the ENSCI111 course, and an increase in individual performance by students on the non-collaborative midterm and final exams in GEOL25 compared to the preceding collaborative tests. The method of grading the collaborative exams in the ENSCI111 course is complex, but provides an incentive ("jackpot") for students to collaborate intelligently without being misled. However, the simpler collaborative approach used subsequently in

GEOL25 tests appears to confer similar learning benefits with lower stakes for the students because the course grade is less dependent on collaborative testing.

Collaborative testing also appears to provide a stabilizing effect on the distribution of grades, allowing test questions to be fewer, more complex and evaluate learning at a higher level on the Bloom taxonomy scale (Bloom, 1956). The additional strategies implemented in these courses to emphasize graphical information during lecture and encourage class participation as well as attendance also appear to contribute in a synergistic way to the overall learning environment. It is hoped that the combination of methods and results described here will be of use to instructors seeking to incorporate them into their geoscience courses.

Acknowledgments

References

Bloom, B.S., 1956, Taxonomy of Educational Objectives, Handbook I: the Cognitive Domain: New York, David McKay Co. Inc.

Cortright, R.N., Collins, H.L., Rodenbaugh, D.W., and DiCarlo, S.E., 2003, Student retention of course content is improved by collaborative-group testing: Advances In Physiology Education, v. 27, p. 102-108.

Felder, R.M., 1993, Reaching the second tier: learning and teaching styles in college science education: Journal of College Science Teaching, v. 23, p. 286-290.

Felder, R.M., and Spurlin, J., 2005, Applications, reliability and validity of the Index of Learning Styles: International Journal of Engineering Education, v. 21, p. 103-112.

Gonzales, P., Guzman, J.C., Partelow, L., Pahlke, E., Jocelyn, L., Kastberg, D., and Williams, T., 2004, Highlights from the Trends in International Mathematics and Science Study (TIMSS) 2003 (NCES 2005-005): Washington D.C., U.S. Department of Education National Center for Education Statistics, U.S. Government Printing Office.

Macdonald, R.H., and Korinek, L., 1995, Cooperative-learning activities in large entry-level geology courses: Journal of Geological Education, v. 43, p. 341-345.

McConnell, D.A., Steer, D.N., and Owens, K.D., 2003, Assessment and active learning strategies in introductory geology courses: Journal of Geoscience Education, v. 51, p. 205.

McConnell, D.A., Steer, D.N., Owens, K.D., Knott, J.R., Van Horn, S., Borowski, W., Dick, J., Foos, A., Malone, M., McGrew, H., Greer, H., and Heaney, P.J., 2006, Using Conceptests to assess and improve student conceptual understanding in introductory geoscience courses: Journal of Geoscience Education, v. 54, p. 61-68.

Moore, R., 2003, Helping students succeed in introductory biology classes: does improving student attendance also improve their grades? Bioscene, v. 29, p. 17-25.

National Academy of Science, 2005, Rising above the gathering storm: energizing and employing America to a brighter economic future: Washington D.C., National Academies Press.

National Research Council, 1996, National Science Education Standards: Washington D.C., National Academies Press - National Committee on Science Education Standards and Assessment, 272 p.

National Science Board, 2007, National Action Plan for addressing the critical needs of the U.S. Science, Technology, Engineering and Mathematics education system: Washington D.Cs, National Science Foundation, 100 p.

Rao, S.P., Collins, H.L., and DiCarlo, S.E., 2002, Collaborative testing enhances student learning: Advances In Physiology Education, v. 26, p. 37-41.

Reynolds, S.J., and Peacock, S.M., 1998, Landscape appreciation and critical thinking in introductory geology courses: Journal of Geoscience Education, v. 46, p. 421-426.

Schwab, B.E., 2005, An assessment of multiple-choice cooperative exams in 'earthquake country'; a large introductory Earth science class, in Hemphill-Haley, M.A., ed., Abstracts with Programs - Geological Society of America, Volume 37: United States, Geological Society of America (GSA): Boulder, CO, United States, p. 263.

Shen, J., Hiltz, S.R., and Bieber, M., 2008, Learning strategies in online collaborative examinations: IEEE Transactions on Professional Communication, v. 51, p. 63-78.

Shindler, J.V., 2004, "Greater than the sum of the parts?" Examining the soundness of collaborative exams in teacher education courses: Innovative Higher Education, v. 28, p. 273-283.

Yuretich, R.F., Khan, S.A., Ledkie, R.M., and Clement, J.J., 2001, Active-learning methods to improve student performance and scientific interest in a large introductory oceanography class: Journal of Geoscience Education, v. 49, p. 111-119.

Zipp, J.F., 2007, Learning by Exams: The Impact of Two-Stage Cooperative Tests: Teaching Sociology, v. 35, p. 62.