CBE—Life Sciences Education
Vol. 11, 392–401, Winter 2012

Article

Collaborative Testing Improves Performance but Not Content Retention in a Large-Enrollment Introductory Biology Class

Hayley Leight, Cheston Saunders, Robin Calkins, and Michelle Withers

Department of Biology, West Virginia University, Morgantown, WV 26506

Submitted April 16, 2012; Revised August 8, 2012; Accepted August 10, 2012
Monitoring Editor: Diane Ebert-May

Collaborative testing has been shown to improve performance but not always content retention. In this study, we investigated whether collaborative testing could improve both performance and content retention in a large, introductory biology course. Students were semirandomly divided into two groups based on their performances on exam 1. Each group contained equal numbers of students scoring in each grade category ("A"–"F") on exam 1. All students completed each of the four exams of the semester as individuals. For exam 2, one group took the exam a second time in small groups immediately following the individually administered test. The other group followed this same format for exam 3. Individual and group exam scores were compared to determine differences in performance. All but exam 1 contained a subset of cumulative questions from the previous exam. Performances on the cumulative questions for exams 3 and 4 were compared for the two groups to determine whether there were significant differences in content retention. Even though group test scores were significantly higher than individual test scores, students who participated in collaborative testing performed no differently on cumulative questions than students who took the previous exam as individuals.

INTRODUCTION

At large research universities, it is not uncommon for introductory science courses to have enrollments of 200 or more students (Smith *et al.*, 2005). Based on the grading time necessary for such large numbers of students, examinations for these classes tend to be made up primarily of multiple-choice questions and in this context, provide an objective, time-efficient method for evaluating student performance (Straits and Gomez-Zwiep, 2009). Computer-based testing offers a convenient vehicle for administering multiple-choice exams to large numbers of students and as such is becoming increasingly commonplace in college classrooms (Clariana and Wallace, 2002). Computer-based exams have many attractive features. For example, they facilitate standard testing procedures; allow for accurate, objective scoring; provide a mechanism for quantitative assessment of student learning; and permit the assessment of cognitive and perceptual performances of students, as well as of their content knowledge (Mead and Drasgow, 1993; Rosenfeld *et al.*, 1993; de Beer and Visser, 1998; Metz, 2009). An additional benefit of computer-based exams is the ability to administer them outside class at a variety of times. This offers students the flexibility of scheduling their examinations at times that best fit both their personal preferences and busy schedules, which may lead to reductions in exam anxiety and in the number of students who miss exams (Stowell and Bennett, 2010).

One weakness of computer-based exams can be the lack of postexamination feedback for students, as they typically receive only a numerical grade. When tests are offered at multiple times during an exam window, several equivalent versions of each multiple-choice question are generated to reduce the probability that students testing at later times will receive the exact combination of questions as students testing

DOI: 10.1187/cbe.12-04-0048
Address correspondence to: Michelle Withers (Michelle.Withers@mail.wvu.edu).

earlier. Unfortunately, this drastically increases the requisite size of question banks for computer-based tests. Due to their large size, these question banks are time-consuming to create and, therefore, not openly shared with students. As a result, students must review their exams in a monitored environment, which may reduce the number of students who opt to do so. Without postexamination feedback, the exams are primarily tools of evaluation and miss an opportunity to facilitate learning.

Assessments such as exams are best used as tools to help instructors better understand the relationship between what we teach and what students learn (Tanner and Allen, 2004) and to help students improve retention and comprehension of content. Testing as a study strategy has been shown to improve content retention due to repeated recall efforts, a phenomenon referred to as the "testing effect" (Roediger and Karpicke, 2006). Multiple-choice exams can also invoke the testing effect, resulting in improvements on subsequent exams (Marsh and Roediger, 2007). However, when students receive a numerical grade only, exams serve primarily as mechanisms for evaluation and have little impact on student learning and content retention (Epstein *et al.*, 2001, 2002). Furthermore, some students succumb to a phenomenon known as the "negative testing effect," in which their recollection of incorrect choices interferes with the learning of correct content (Roediger and Marsh, 2005). Rao *et al.* (2002) demonstrated that incorporating examination formats that allow students opportunities to receive feedback on mistakes made on multiple-choice questions may reduce or preclude this negative impact on learning.

Group testing is a promising way to bring the power of collaborative learning to bear on the discussion and analysis of exam questions after students have completed an exam once as individuals (Millis and Cottell, 1998; Michaelson *et al.*, 2002; Hodges, 2004). Incorporating group exams into the computer-based testing format could provide a time-efficient mechanism to boost the learning potential of these exams. Collaborative testing improves performance (Stearns, 1996; Sumangala *et al.*, 2002; Giuliodori *et al.*, 2008; Eaton, 2009; Haberyan and Barnett, 2010) and motivation (Hodges, 2004; Kapitanoff, 2009), decreases test anxiety (Muir and Tracy, 1999, Zimbardo *et al.*, 2003; Hodges, 2004; Kapitanoff, 2009), and effectively evaluates student learning (Russo and Warren, 1999). It is also viewed positively by students (Cortright *et al.*, 2003; Lusk and Conklin, 2003; Mitchell and Melton, 2003; Zimbardo *et al.*, 2003; Shindler, 2004; Woody *et al.*, 2008; Sandahl, 2010). While studies consistently have demonstrated improvements in student performance on collaborative exams, the ability of collaborative testing to improve content retention is still in question. Some studies report an improvement in content retention from collaborative testing (Rao *et al.*, 2002; Cortright *et al.*, 2003; Bloom, 2009), while others show no effect (Lusk and Conklin, 2003; Woody *et al.*, 2008; Sandahl, 2010). Given the inconsistency of these findings and the extra time and resources required to add collaborative testing into an existing examination format, we wanted to know whether collaborative testing would indeed improve student learning in an introductory science class. In this study, we investigated whether collaborative examinations can improve both performance and content retention when added to a computer-based testing format for a large-enrollment introductory biology course.

METHODS

Course Context

Biology 115: Principles of Biology, is a first-semester, introductory-level course with a laboratory that introduces students to basic concepts in cellular, molecular, and evolutionary biology and fundamental science process skills. It is the first of a five-course series required for biology majors and serves as a specific requirement for several undergraduate science degrees on campus. In addition, this course fills a General Education Curriculum requirement for non–science majors at West Virginia University, Morgantown. The course consists primarily of freshmen seeking degrees in biology, chemistry, or life sciences–related disciplines. A very small proportion of the students are seeking degrees in other science and non–science disciplines. The class is roughly split between males and females.

Course Structure

To examine the impact of collaborative testing on student learning, we used a single section (~250 students) of Biology 115: Principles of Biology, during the Fall semester of 2010. The section employed an active-learning format and was taught by a discipline-based education researcher trained in scientific teaching by the National Academies Summer Institute on Undergraduate Biology Education who has taught the course since 2007. Group-learning activities, such as personal response system (clicker) questions, case studies, discussion, and problem solving were employed on a daily basis to engage students with the course material. In addition to the lecture-based component of the course, all students were enrolled in an accompanying laboratory section. Final grades were determined from five course examinations, concept inventory pre- and posttests, formative assessments, and laboratory exercises. The objective course examinations were computer-based and consisted of multiple-choice, multiple-correct, true/false, and sequencing problems.

Research Design

To evaluate the effect of collaborative testing on content retention in this large-enrollment introductory biology course, we employed a randomized cross-over design (Cortright *et al.*, 2003; Sandahl, 2010). We elected to use a randomized cross-over design due to its unique characteristic: each subject has the ability to serve as his or her own control (Rietbergen and Moerbeek, 2011). Essentially, by randomly splitting the class in two, we were able to run the experiment twice during the semester with each group serving once as the experimental group and once as the control group that controlled for coincidental differences in the two samples. Based on scores from exam 1, students in each grade category ("A"–"F") were randomly assigned to one of two equally sized groups (A or B). Due either to withdrawal or nonparticipation in the group exams, the group sizes for A and B were 92 and 104 students, respectively, at the conclusion of this study. Power analysis using the effect size of 0.06 calculated from data reported by Cortright *et al.* (2003), a power of 0.8, and a two-tailed alpha of 0.05, yielded a sample size requirement of 90.

All students, regardless of group designation, completed each course exam as individuals in the Biology Department

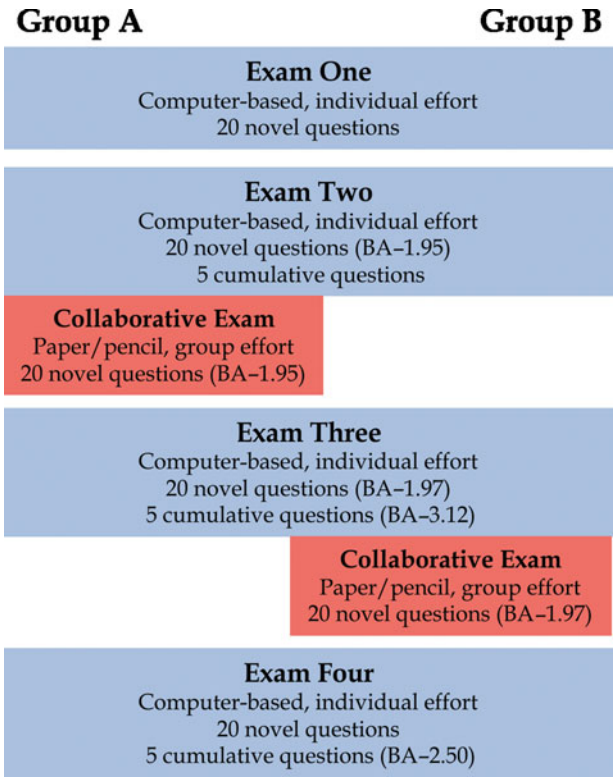

Figure 1. Research design. All students took exams 1–4 as individuals. Group A took exam 2 and group B took exam 3 in small groups immediately following the individual exam. Each exam of the semester occurred 3 wk after the previous exam. Bloom's averages are reported for the novel and cumulative questions involved in the study only. BA, Bloom's average (level 1: knowledge; level 2: comprehension; level 3: application; level 4: analysis; level 5: synthesis; and level 6: evaluation).

computer-testing facility. Computer-based exams were administered outside regular class time over the period of 1 wk (Monday to Friday). Students scheduled their own exam times during that week to suit their preferences and schedules. To keep students who tested later in the week from receiving the exact same exam questions as students who tested earlier in the week, we created computer-based test banks with multiple versions of each question. Alternate questions were similar in topic and cognitive level. For a given exam, each student was tested on the same novel and cumulative topics.

In addition to the exams administered to individuals (referred to as *individual exams* in this paper), students in group A took exam 2 and students in group B took exam 3 in small groups immediately following their individual computer-based exams (Figure 1). Group exam periods followed each of the individual, computer-based exam periods, allowing students taking the group exam the same exam scheduling flexibility as other students. Students participating in the group format were relocated by a graduate teaching assistant to a second room immediately following the individual exam period, and they completed a paper version of the exam in groups. For collaborative groups to be effective, they must be small in size so that each student has an opportunity to participate (Cohen, 1994); therefore, we instructed our students to work in groups of two to four students. Students self-assembled into groups and, thus, may or may not have known the other members of their group. Students were instructed to discuss each question and arrive at a consensus on each answer. Students were allotted the same amount of time to complete the individual and group exams. At the end of the group exam period, each group submitted a single group exam for evaluation. To encourage a good faith effort on the collaborative exams, we added percentage points to students' individual exam scores based upon the following scheme: group exam scores of 90–100%, 80–89%, and 70–79% resulted in increases in individual exam scores of 5%, 3%, and 1%, respectively. For example, a student who scored 85% on the individual exam and 95% on the group exam received a total score of 90%. There was no increase in individual grades for group exam scores below 70%. Because the total exam score could not be less than the individual score, students who opted not to participate in the group exam did not suffer any penalty to their individual grades. Exam scores used for analysis in this study did not contain the additional incentive points from the group exams.

The individual exams (2, 3, and 4) completed for this study consisted of 20 novel questions covering new material and five cumulative questions covering material from the previous exam, that is, the five cumulative questions on exam 3 are a subset of the 20 novel questions included in the individual and group versions of exam 2 (Figure 1). The set of cumulative questions served as the basis for gauging students' retention of previous material. Cumulative questions were omitted from the group exams to guarantee complete separation of the topics the two groups experienced in a collaborative manner; that is, group A only had new material from exam 2 and group B only had new material from exam 3 on their respective group exams. To determine whether the exams presented similar cognitive challenges, we calculated average Bloom's scores for the novel exam questions from exams 2 and 3 and the cumulative questions from exams 3 and 4 using the Blooming Biology Tool (Crowe *et al.*, 2008). To determine whether individual performances differed from those of the groups, we compared average individual scores with average group scores for the same exam. In an effort to examine differences in students' retention of previously tested material, we compared the performances of groups A and B on the cumulative questions of exams 3 and 4. All statistical analyses were performed using the Statistical Package for Social Sciences (SPSS) software. Variables are reported as mean $\pm$ SE.

During the last week of class, students completed an online survey (modified from Cortright *et al.*, 2003; Table 1) to evaluate their perceptions of the collaborative testing format. Students were asked to rate their level of agreement with 13 statements about the collaborative testing process using a Likert scale. Students received credit for completing the survey equivalent to an online reading preparation quiz for the class. Students were allowed to drop a subset of these quizzes, which included the survey, without penalty. The majority of students who participated in collaborative testing (83.2%) completed the survey.

Table 1. Student perceptions of collaborative testing[a]

Statement in survey	Student response (mean ± SE)
1. The purpose of and rationale behind the collaborative testing process was fully explained.	4.02 ± 0.08
2. The collaborative testing process was not too lengthy or complex in its format.	3.99 ± 0.08
3. The peer discussions on the group testing improved my level of confidence on the answers.	3.64 ± 0.10
4. Every group member contributed to the learning process during the pyramid exam.	3.77 ± 0.09
5. The level of peer (group) discussions was very high.	3.79 ± 0.08
6. The immediate feedback given by the peer discussions was very positive.	3.64 ± 0.08
7. The level of peer (group) discussions enhanced my understanding of the concepts.	3.42 ± 0.08
8. My level of involvement during the collaborative exam discussions was high.	4.27 ± 0.06
9. I was able to later recall concepts because I had the opportunity to previously discuss them within the group.	3.50 ± 0.09
10. This testing methodology provided the opportunity to discuss incorrect answers and fill in knowledge gaps.	3.80 ± 0.08
11. This testing methodology was educationally attractive due to the novelty of this style and format.	3.65 ± 0.09
12. This testing methodology was less stressful than traditional testing methods.	3.90 ± 0.09
13. I would be interested in further classes with similar group testing methodologies.	3.90 ± 0.09

[a]Students responded using a Likert scale (1: completely disagree; 2: disagree; 3: neither agree nor disagree; 4: agree; 5: completely agree) to indicate their level of agreement with the 13 statements shown in the table. The majority of students (83.2%) participating in collaborative testing responded to the survey.

RESULTS

Students Performed Better on Collaborative Exams Than on Individual Exams

To determine whether collaborative testing improved student performance on examinations, we compared individual exam scores with group exam scores (Figure 2). For exams 2 and 3, mean group scores were significantly higher than mean individual scores ($p < 0.001$). For exam 2, the mean group score (74.5 ± 1.63%) was more than 10% higher than the mean individual score for students in group A (61.9 ± 1.89%). We found a similar trend for exam 3, for which the mean group score (84.1 ± 1.13%) was more than 15% higher than the mean individual exam score for students in group B (63.7 ± 1.64%). To determine whether the increases in mean group scores were due simply to elevating the performances of lower-scoring students up to but not above the level of the highest-scoring students, we compared each student's individual score with his or her group score (Figure 3). The vast majority (93.9%) of students scored higher on their group exams than on their individual exams.

Collaborative Examinations Did Not Increase Retention of Course Material

To examine the effect of collaborative testing on content retention, we compared the responses to cumulative questions by groups A and B for exams 3 and 4 (Figure 4). The average number of correct responses to cumulative questions on exam 3 was not significantly different ($p = 0.166$) between group A (2.36 ± 0.16), who took the previous exam collaboratively, and group B (2.09 ± 0.12), who took the previous exam individually. On exam 4, we reversed the roles of the two groups, such that group B experienced the previous exam collaboratively, and saw a similar result, in which the average number of correct responses to cumulative questions was not

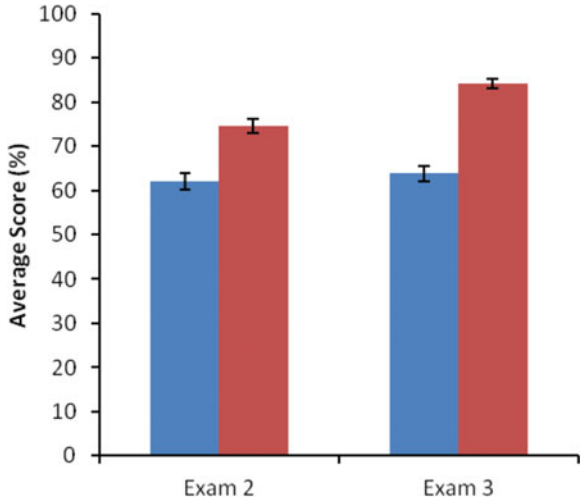

Figure 2. Comparison of group and individual exam performances. Average group exam scores (red bars) were significantly higher ($p < 0.001$, Wilcoxon matched-pairs signed-ranks test) than average individual scores (blue bars).

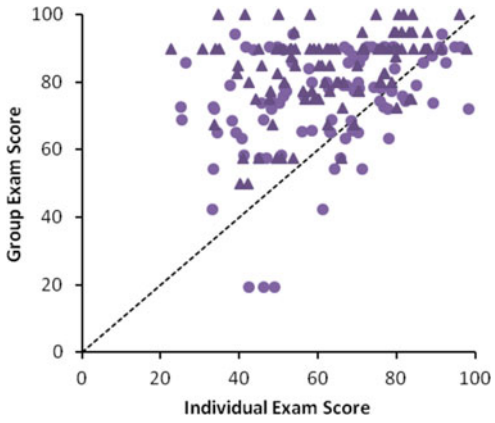

Figure 3. Group exam scores plotted as a function of individual scores. Group A, exam 2 (circles) and group B, exam 3 (triangles) percentages. The majority of data points (93.9%) fall above the dotted line ($y = x$).

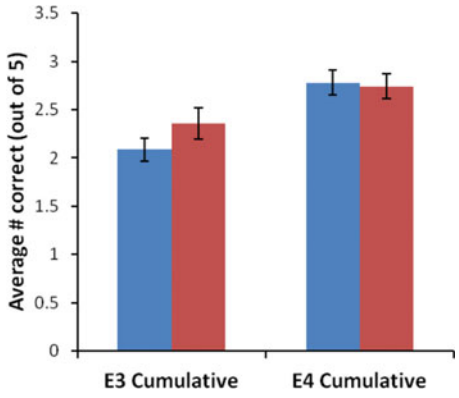

Figure 4. Comparison of the performances by group and individual testers on cumulative questions. Average correct responses on cumulative questions by students who took the previous exam collaboratively (red bars) was not significantly higher than for students who took the previous exam as individuals (blue bars; $p = 0.166$ and 0.846, respectively, two-tailed t test).

significantly different between the two groups (B = 2.74 ± 0.13; A = 2.78 ± 0.13; $p = 0.846$).

In the initial comparison of performances by the two groups on cumulative material, we clustered all students together in their respective groups regardless of whether their group scores were higher or lower than their individual performances. We were curious to know whether the relationship between the students' individual and group performances (i.e., whether their group scores were higher or lower than their individual scores) had any effect on their content retention. Therefore, we separated group A for exam 3 and group B for exam 4 into students whose group scores increased and students whose group scores decreased, and compared their performances on the cumulative questions with those of students who took the previous exam individually (Figure 5). For exam 3, there was a significant effect on the average number of correct responses to cumulative questions between the three groups ($p = 0.005$). Specifically, students whose group scores decreased correctly answered more cumulative questions (3 ± 0.3) than did students whose group scores increased (2.09 ± 0.17, $p = 0.011$) or students who took the previous exam individually (2.09 ± 0.12, $p = 0.006$). A similar trend was observed for exam 4, but the effect was not significant across the three groups ($p = 0.313$). The average cumulative score on exam 4 for students whose collaborative score decreased (3.43 ±0.48) was not significantly higher than either students whose collaborative score increased (2.69 ± 0.13, $p = 0.393$) or students who took the previous exam individually (2.78 ± 0.13, $p = 0.549$).

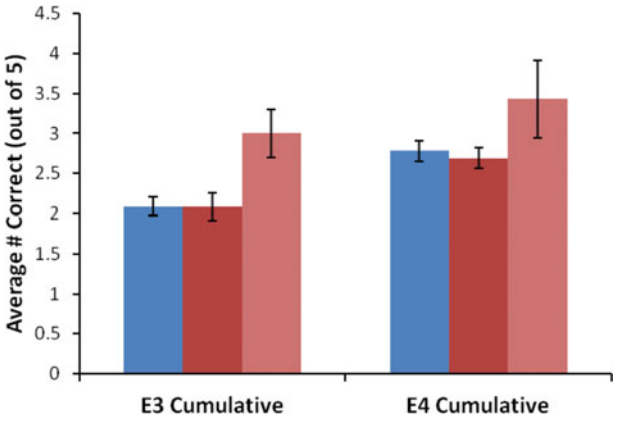

Figure 5. Comparison of cumulative question scores by students whose group scores either increased or decreased. Students whose group scores decreased (light red bars) averaged a significantly higher cumulative score on exam 3 (E3) than students whose group score increased (dark red bars, $p = 0.011$, simple analysis of variance [ANOVA] using Bonferroni post hoc) and students who took the previous exam as individuals (blue bars, $p = 0.006$). There was no significant difference between the cumulative scores of the three groups on exam 4 (E4, $p = 0.313$, simple ANOVA using Bonferroni post hoc).

Individual Performances Are Positively Correlated with Content Retention

Students whose group scores decreased tended to perform better on cumulative questions and the majority of students (68.8%) whose group scores decreased scored at or above a middle "C" on their individual exams (Figure 3). Therefore, we wanted to determine the relationship between individual performance and content retention (Figures 6 and 7). First, we examined the relationship between individual scores and performance on cumulative questions on the subsequent exam (Figure 6). We found a moderate positive correlation between individual exam scores and subsequent cumulative scores on both exam 3 (Figure 6A, $r = 0.571$, $p < 0.001$) and exam 4 (Figure 6B, $r = 0.546$, $p < 0.001$) regardless of whether students took the previous exam individually or collaboratively.

Given the positive correlation between individual scores and subsequent performance on cumulative questions, we were curious as to whether the collaborative testing format had different effects on high- and low-performing students (Figure 7). Therefore, we separated the two groups based on their individual exam grades ("A"–"F") and compared their performances on cumulative questions for exam 3 (Figure 7A) and exam 4 (Figure 7B). While there was a statistically significant relationship between letter grade and the number of correct responses on cumulative questions on subsequent exams for both exams ($p < 0.001$), there were no significant differences in performance on cumulative questions between students who took the previous exams individually or collaboratively at any letter grade for either exam 3 ($F = 0.038$, $p = 0.847$) or exam 4 ($F = 0.173$, $p = 0.678$).

Students Responded Positively to the Collaborative Testing Format

Finally, a survey instrument was used to evaluate student perception of and participation in the collaborative testing methodology. The questionnaire evaluated student reaction to the purpose and length of the testing process, the level of involvement of group members during the discussion, as well as the students' attitudes and perceptions on the testing process (Table 1). Students reported a high level of involvement for all members within their groups (questions 5 and 8), as well as an enhanced understanding of concepts and increase in confidence in answers as a result of peer discussions (questions 3 and 7). Overall, students reported that they enjoyed the novelty of the testing methodology (question 11), found collaborative exams to be less stressful than a traditional

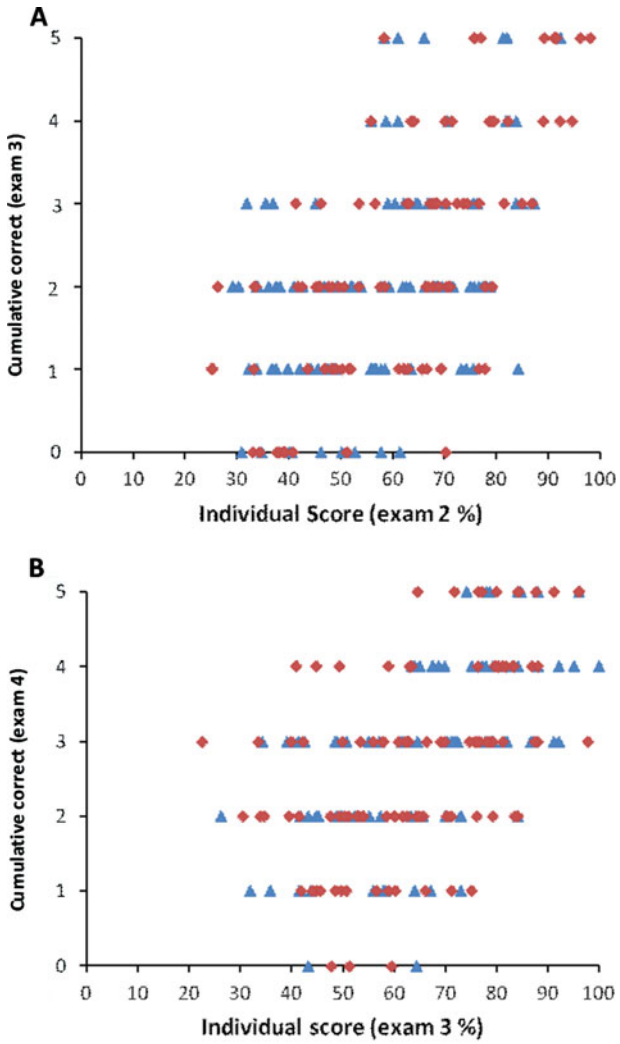

Figure 6. Cumulative question scores as a function of individual exam scores. Individual exam scores are positively correlated with performances on cumulative questions for (A) exam 3 ($r = 0.571$, $p < 0.001$, Pearson's rho correlation) and (B) exam 4 ($r = 0.546$, $p < 0.001$) regardless of whether the previous test was taken individually (blue triangles) or as a group (red diamonds). Students who took the previous exam as individuals (exam 3, $r = 0.446$, $p < 0.001$; exam 4, $r = 0.593$, $p < 0.001$) show the same positive correlation between individual and cumulative score as students who took the previous exam collaboratively (exam 3, $r = 0.666$, $p < 0.001$; exam 4, $r = 0.506$, $p < 0.001$).

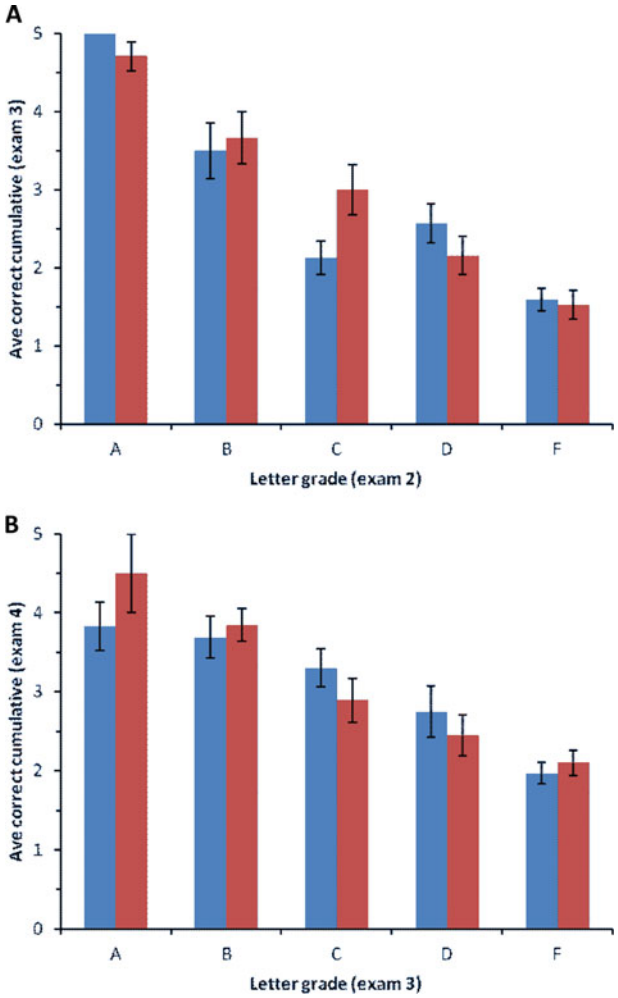

Figure 7. Comparison of cumulative question scores by students scoring in different grade categories on individual exams. There was a statistically significant relationship between a student's letter grade and the number of correct responses on subsequent exams for both (A) exam 3 and (B) exam 4 ($p < 0.001$, 2×5 factorial ANOVA), regardless of whether students took the previous exam individually (blue bars) or collaboratively (red bars). The average number of correct cumulative responses was not significantly different between the individual and collaborative groups for either exam (E3, $p = 0.847$; E4, $p = 0.678$).

exam format (question 12), and would be interested in future classrooms with similar testing pedagogies (question 13).

DISCUSSION

Our data support the previous finding that students perform better on collaborative tests than on individual tests (Stearns, 1996; Sumangala *et al.*, 2002; Cortright *et al.*, 2003; Lusk and Conklin, 2003; Giuliodori *et al.*, 2008; Woody *et al.*, 2008; Bloom, 2009; Eaton, 2009; Haberyan and Barnett, 2010; Sandahl, 2010). For both trials, mean exam scores were significantly higher when exams were completed by a group than when completed by individuals. Additionally, the majority of the group scores were higher than the individual scores, demonstrating that most groups scored higher than the top individual in each group. This suggests that students were working together to pool their knowledge and understanding rather than relying on the top student to provide all of the answers (Giuliodori *et al.*, 2008) and is in agreement with their perception that the majority of group members contributed equally to the collaborative testing process. Because learning is fostered by feedback, students should get more benefit from collaborative exams when they interact with classmates as a "community of learners," promoting the "elaboration of knowledge structures" and fostering individual awareness of personal learning processes (Wood, 2009). In addition, collaborative testing provides opportunities for students to

Table 2. Comparison of collaborative testing studies[a]

	Improved content retention	Study population	Course content	Collaborative format: individual exam followed by group exam	Class format[b]
Our study	No	Science majors, primarily freshmen	Introductory Biology	Yes	Active, collaborative learning
Bloom, 2009	Yes	Theater nonmajors, primarily freshmen and sophomores	Introduction to Theater	Yes (access to books and notes during collaborative exam)	NR
Cortright *et al.*, 2003	Yes	Science majors, primarily juniors and seniors	Physiology of Exercise	Yes	Lecture
Lusk and Conklin, 2003	No	Nursing students, sophomores	Fundamentals of Nursing	Yes	NR
Sandahl, 2010	No	Nursing students, seniors	Medical–Surgical Nursing II	No (no initial individual component)	NR
Woody *et al.*, 2008	No	Psychology majors, seniors	Theories of Counseling & Psychology of Prejudice	No (no initial individual component)	NR

[a]Results, methods, and design parameters for studies investigating the impact of collaborative testing on content retention.
[b]NR, not reported.

evaluate personal understanding of course material without the stress of a high-stakes assessment (Cortright *et al.*, 2003; Zimbardo *et al.*, 2003; Shindler, 2004). Our study demonstrates that using collaborative testing enhances performance.

Although our group exam scores were higher than the individual exam scores, content retention was no better for students who completed the previous test collaboratively. This finding is in agreement with several studies that also demonstrated no significant impact of collaborative testing on content retention (Lusk and Conklin, 2003; Woody *et al.*, 2008; Sandahl, 2010), but is in contrast to others (Bloom, 2009; Cortright *et al.*, 2003; see Table 2). It is unlikely that the failure to observe an improvement in content retention by collaborative testing was due to a general failure of our students to engage in the group testing process, since improvement in group scores over individual scores for our study (14.8%) was comparable to the two studies that did report increased content retention (16% [Bloom, 2009] and 18% [Cortright *et al.*, 2003]). In addition, students in our study were accustomed to the group work format through participation in daily group-learning activities in class. Based on the end-of-term perception survey, students also tended to agree that they and their fellow classmates were engaged in the group testing process and that the group discussions aided their understanding. There are differences in other aspects of these studies, such as study population and course content, which could contribute to the different findings.

Our student population is most similar to Bloom's (2009) population in grade level, although the course content in these two studies is the most dissimilar. Both studies investigate collaborative testing in introductory courses with large freshmen populations. However, the Bloom (2009) study examines an introductory theater course for nonmajors, in which "the majority of the questions on the exams focused on information recall, and a few required the application of concepts covered in the class." This is in contrast to our introductory biology course for science majors, which addresses the complex processes and concepts underlying cellular and molecular biology, genetics, and evolution. The average Bloom's level was approximately 3, application level, for the cumulative questions on exams 3 and 4 that were used to evaluate content retention. Differences in difficulty and/or the cognitive levels of material covered by the two courses may contribute to the differences seen in content retention. For example, the average number of students who scored 100% on the group exam in Bloom's study was 56% compared with 0% and 9% of students in groups A and B in our study, respectively, who scored 100% on the group exams (Figure 3). This suggests that the level of difficulty was higher for the material evaluated by our introductory biology exams; this could have an impact on the ability for collaborative testing to significantly improve retention of this material.

Studies that have measured improved content retention from repeated testing, the testing effect, have measured recall of simple information (Roediger and Karpicke, 2006). Perhaps one mechanism by which collaborative testing improves content retention of lower-cognitive-level material is a combination of repeated recall, which leads to the testing effect, combined with opportunities to correct mistakes and thus avoid the negative testing effect, in which students recall incorrect choices (Roediger and Marsh, 2005). This could result in improved recall without improved understanding, that is, remembering that the answer associated with choice "a" is correct, without understanding why that is the case, especially if retention is being tested with the same, rather than equivalent, questions. In our study, all students in groups A and B took the same group versions of exams 2 and 3, respectively. However, when they took the subsequent exam, the cumulative questions with which retention was tested could be equivalent, that is, similar in cognitive level and conceptual content, but not necessarily exactly the same. If our students only remembered that certain answers were associated with certain questions from the group exam, then they would not

be any more likely to answer an equivalent question correctly than someone who took the previous exam as an individual.

A second difference between our and Bloom's (2009) studies that may contribute to the observed differences in retention is the format for the collaborative exam. In both studies students completed the individual test, which was followed immediately by the group test. However, unlike our study, in which students only consulted one another during the group test, students in the Bloom study were allowed to consult their books and notes in addition to their peers during the collaborative test. This may have provided more opportunity to correct wrong answers than when just relying on peers. This might explain, in part, the high percentage of perfect scores on the collaborative exams for that study. If increasing the chances for students to correct mistakes by consulting books and notes leads to a reduction in the negative testing effect, as demonstrated by Rao *et al.* (2002), then the improvement in content retention observed in the Bloom study could result, in part, from this practice.

The content in our introductory biology course was much more similar to that of Cortright *et al.* (2003). Both courses address biology-related topics; however, our student populations are at very different levels of expertise within their majors. Participants in that study were third- and fourth-year majors in exercise and sport science who were taking a required course entitled the Physiology of Exercise. In contrast, participants in our study were primarily freshmen taking an introductory biology course for science majors; for many, this may have been their first college-level science course. Cortright *et al.* (2003) argue that student retention of course content is short-lived, regardless of whether students are "experienced" or "naïve." Yet educational literature suggests that learning in a discipline progresses from novice to expert (Benner, 1984; Donovan *et al.*, 1999; Wood, 2009). One can, therefore, logically assert that students who are juniors and seniors in a discipline have at least a beginning sense of a "coherent structure of knowledge" in contrast to that of the novice, whose disciplinary understanding has been described as a "collection of unrelated facts which are difficult to memorize and retain" (Wood, 2009).

A plausible explanation for the differences in content retention observed between our study and that of Cortright *et al.* (2003) is that novice and veteran students undergo unique learning experiences during collaborative testing. Experienced students may be able to fit the new concepts into their pre-existing knowledge structure, thus facilitating retention, whereas naïve students may find this information only contributes to their collection of unrelated facts that are hard to remember. In addition, expert students are likely to present more cogent explanations for why an answer is right or wrong, thus helping their peers better understand and remember the information. The confidence that comes with increasing expertise is also likely to play a role in the way that collaborative testing affects learning by freshmen versus upperclassmen. This seems to be supported by the fact that the majority of students in our study whose group scores decreased scored above 75% on the individual exam. This suggests that even higher-achieving students could be convinced by others to change a previously correct answer to an incorrect one.

Interestingly, students whose group scores decreased demonstrated better content retention. There are two possible explanations for this finding. First, researchers have recently found that initially getting a question wrong helps students better recall information later (Kornell *et al.*, 2009). Students who were talked out of a right answer may have been more likely to check that material again after the group test and therefore may have remembered the correction. Alternatively, students whose group grades decreased generally had higher individual grades and higher individual grades were positively correlated with greater retention. Regardless of the testing format, there was a positive correlation between individual performance and content retention. Additionally, when sorted by exam grade ("A"–"F"), there was still no impact of collaborative testing on content retention for students at different achievement levels. These results underscore the important relationship between individual exam scores and performance on the cumulative questions on subsequent exams. While not surprising, these data reinforce the relevance of individual exam scores as vital indicators of future performance.

A second aspect that may have contributed to the differences in the results of our and the Cortright *et al.* (2003) studies is the exam format. Students in both studies took the collaborative exams in small groups without any additional resources, such as books or notes. However, in our study, the collaborative exam consisted of more questions. For the collaborative exam, our students retook the 20 novel questions from the individual exam. Then, on the subsequent exam, students answered a subset of those 20 questions to test for retention of content. In the study by Cortright *et al.* (2003), the collaborative exam was already reduced to a subset of the original exam, and students were then tested for retention with the exact subset of questions. This could contribute to the observed differences in retention in two ways. First, having more questions on the collaborative exam for the freshman course could exacerbate the aforementioned differences in how novices and experts deal with information. The increased number of questions that the freshmen faced on the collaborative exam could have presented a cognitive overload that contributed to a perception of the material as a "collection of unrelated facts" and reduced their ability to retain the information. Secondly, testing retention with the exact set of questions from the collaborative exam could have resulted in improved recall without improved understanding, as mentioned earlier for the Bloom (2009) study. If that is the phenomenon being promoted by collaborative testing, then our method of testing for retention using equivalent but not necessarily exact questions would fail to measure an improvement.

A final element that may contribute to the differences seen in content retention between our study and that of Cortright *et al.* (2003) is the method of instruction employed. Cortright *et al.* (2003) used lectures to instruct their upper-level physiology course, while the introductory biology course in our study was taught primarily through active learning. Students in our introductory course take preclass quizzes to drive acquisition of fact-based content, which leaves the majority of class time for active learning driven by such activities as clicker questions, think–pair–share, discussion, and case scenarios to help students apply their understanding and practice critical thinking and problem solving. A number of studies have demonstrated improved learning in courses in which students are active rather than passive participants

(Beichner and Saul, 2003; Knight and Wood, 2005; Freeman *et al.*, 2007; Walker *et al.*, 2008). Perhaps it is more difficult for interventions such as collaborative testing to stimulate significant increases in content retention for students who have already experienced the improved learning afforded by reform-based pedagogies, as have the students in our study. Students in a lecture-based class may have more untapped potential for learning improvements by interventions such as group testing.

As with most other studies (Cortright *et al.*, 2003; Lusk and Conklin, 2003; Mitchell and Melton, 2003; Zimbardo *et al.*, 2003; Shindler, 2004; Woody *et al.*, 2008; Sandahl, 2010), students in our study reported favorable views of this testing method, as well as its utility as a teaching and learning tool. Respondents indicated that collaborative testing was a nonthreatening approach to learning and that peer interactions facilitated both their understanding of concepts and their personal confidence in ability to answer questions correctly. There was a high level of agreement that this approach would be desirable in future classes. While students perceived collaborative testing to be both more helpful and less stressful than traditional testing, the absence of improved content retention in this and other studies with varied student populations and content areas (see Table 2) suggests that collaborative testing is not a robust method for improving student learning. Because only two studies have demonstrated improved content retention with collaborative testing, instructors who wish to use this method in their classes should determine, rather than assume, that it improves their students' learning.

REFERENCES

Beichner R, Saul J (2003). Introduction to the SCALE-UP (Student-Centered Activities for Large-Enrollment Undergraduate Programs) project. Proceedings of the International School of Physics, Varenna, Italy, July 2003. www.ncsu.edu/PER/Articles/Varenna_SCALEUP_Paper.pdf (accessed 5 April 2012).

Benner P (1984). From Novice to Expert, Menlo Park, CA: Addison-Wesley.

Bloom D (2009). Collaborative test taking: benefits for learning and retention. Coll Teach *57*, 216–220.

Clariana R, Wallace P (2002). Paper-based versus computer-based assessment: key factors associated with the test mode effect. Br J Educ Technol *33*, 593–602.

Cohen E (1994). Restructuring the classroom: conditions for productive small groups. Rev Educ Res *64*, 1–35.

Cortright R, Collins H, Rodenbaugh D, DiCarlo S (2003). Student retention of course content is improved by collaborative-group testing. Adv Physiol Educ *27*, 102–108.

Crowe A, Dirks C, Wenderoth MP (2008). Biology in Bloom: implementing Bloom's taxonomy to enhance student learning in biology. CBE Life Sci Educ *7*, 368–381.

de Beer M, Visser D (1998). Comparability of the paper-and-pencil and computerized adaptive versions of the General Scholastic Aptitude Test (GSAT) Senior. South African J Psychol *28*, 21–27.

Donovan M, Bransford J, Pellegrino J (1999). How People Learn: Bridging Research and Practice, Committee on Learning, Research and Educational Practice, Washington, DC: National Academy Press.

Eaton T (2009). Engaging students and evaluating learning progress using collaborative exams in introductory courses. J Geosci Educ *57*, 1–10.

Epstein M, Epstein B, Brovsic G (2001). Immediate feedback during academic testing. Psychol Rep *88*, 889–894.

Epstein M, Lazarus A, Calvano T, Matthews K, Hendel R, Epstein B, Brosvic G (2002). Immediate feedback assessment technique promotes learning and corrects inaccurate first responses. Psychol Rec *52*, 187–201.

Freeman S, O'Connor E, Parks J, Cunningham M, Hurley D, Haak D, Dirks C, Wenderoth M (2007). Prescribed active learning increases performance in introductory biology. CBE Life Sci Educ *6*, 132–139.

Giuliodori M, Lujan H, DiCarlo S (2008). Collaborative group testing benefits high- and low-performing students. Adv Physiol Educ *32*, 274–278.

Haberyan A, Barnett J (2010). Collaborative testing and achievement: are two heads really better than one? J Instruct Psychol *37*, 32–41.

Hodges L (2004). Group exams in science courses. New Dir Teach Learn *2004*, 89–93.

Kapitanoff S (2009). Collaborative testing: cognitive and interpersonal processes related to enhanced test performance. Active Learn Higher Educ *10*, 56–70.

Knight J, Wood W (2005). Teaching more by lecturing less. Cell Biol Educ *4*, 298–310.

Kornell N, Hays M, Bjork R (2009). Unsuccessful retrieval attempts enhance subsequent learning. J Exp Psychol Learn Mem Cogn *35*, 989–998.

Lusk M, Conklin L (2003). Collaborative testing to promote learning. J Nurs Educ *42*, 121–124.

Marsh E, Roediger H (2007). The memorial consequences of multiple-choice testing. Psychonomic Bull Rev *14*, 194–199.

Mead A, Drasgow F (1993). Equivalence of computerized and paper-and pencil cognitive ability tests: a meta-analysis. Psychol Bull *114*, 449–458.

Metz A (2009). Quantitative assessment of student learning in large undergraduate science classrooms: approaches and caveats. In: College Science Teachers Guide to Assessment, ed. T Lord, D French, and L Crow, Arlington, VA: NSTA Press, 31–34.

Michaelson L, Knight A, Fink L (eds.) (2002). Team-Based Learning: A Transformative Use of Small Groups, Westport, CT: Praeger.

Millis B, Cottell P (1998). Cooperative Learning for Higher Education Faculty, Phoenix, AZ: American Council on Education and Oryx Press.

Mitchell N, Melton S (2003). Collaborative testing: an innovative approach to test taking. Nurse Educ *28*, 95–97.

Muir S, Tracy D (1999). Collaborative essay testing. Coll Teach *47*, 33–37.

Rao S, Collins H, DiCarlo S (2002). Collaborative testing enhances student learning. Adv Physiol Educ *26*, 37–41.

Rietbergen C, Moerbeek M (2011). The design of cluster randomized crossover trials. J Educ Behav Stat *36*, 472–490.

Roediger H, Karpicke J (2006). Test-enhanced learning: taking memory tests improves long-term retention. Psychol Sci *17*, 249–255.

Roediger H, Marsh E (2005). The positive and negative consequences of multiple-choice testing. J Exp Psychol Learn Mem Cog *31*, 1155–1159.

Rosenfeld P, Booth-Kewley S, Edwards J (1993). Computer-administered surveys in organizational settings: alternatives, advantages, and applications. Am Behav Sci *63*, 485–511.

Russo A, Warren S (1999). Collaborative test taking. Coll Teach *47*, 18–20.

Sandahl S (2010). Collaborative testing as a learning strategy in nursing education. Nurs Educ Perspect *31*, 142–147.

Shindler J (2004). Greater than the sum of the parts? Examining the soundness of collaborative exams in teacher education courses. Innov High Educ *28*, 273–283.

Smith A, Stewart R, Shields P, Hayes-Klosteridis J, Robinson P, Yuan R (2005). Introductory biology courses: a framework to support active learning in large enrollment introductory science courses. Cell Biol Educ *4*, 143–156.

Stearns S (1996). Collaborative exams as learning tools. Coll Teach *44*, 111–112.

Stowell J, Bennett D (2010). Effects of online testing on student exam performance and test anxiety. J Educ Comput Res *42*, 161–171.

Straits W, Gomez-Zwiep S (2009). Better multiple-choice assessments. In: College Science Teachers Guide to Assessment, ed. T Lord, D French, and L Crow, Arlington, VA: NSTA Press, 45–48.

Sumangala R, Collins H, DiCarlo S (2002). Collaborative testing enhances student learning. Adv Physiol Educ *26*, 37–41.

Tanner K, Allen D (2004). Approaches to biology teaching and learning: from assays to assessments—on collecting evidence in science teaching. Cell Biol Educ *3*, 69–74.

Walker J, Cotner S, Baepler P, Decker M (2008). A delicate balance: integrating active learning into a large lecture course. CBE Life Sci Educ *7*, 361–367.

Wood W (2009). Innovations in teaching undergraduate biology and why we need them. Annu Rev Cell Dev Biol *25*, 93–112.

Woody W, Woody L, Bromley S (2008). Anticipated group versus individual examinations: a classroom comparison. Teach Psychol *35*, 13–17.

Zimbardo P, Butler L, Wolfe V (2003). Cooperative college examinations: more gain, less pain. J Exp Educ *71*, 101–125.