# Retention following Two-Stage Collaborative Exams Depends on Timing and Student Performance

**James E. Cooke,[†‡§*] Laura Weir,[†‖¶] and Bridgette Clarkston[‖#]**

[‡]Biology Department and [‖]Carl Wieman Science Education Initiative, University of British Columbia, Vancouver, BC V6T 1Z4, Canada

## ABSTRACT

**Multistage collaborative exams are implemented to enhance learning and retention of course material. However, the effects of multistage collaborative exams on retention of course content are varied. These discrepancies may be due to a number of factors. To date, studies examining collaborative exams and content retention have used questions that all, or mostly, require students to select an answer, rather than generate one of their own. However, content retention can improve when students generate their own responses. Thus, we examined the effect of collaborative exams with open-ended questions on retention of course content. Retention was measured at two time periods; one relatively shortly (9 days) following a collaborative exam and another over a longer time period (23 days). Furthermore, we examined whether content retention differed for low-, mid-, or high-performing students. Our results suggest that collaborative exams offer retention benefits at relatively long time periods between pre- and posttests, but not over shorter time periods. Retention varied across students in different performance categories. Our study, the first to use only open-ended questions, showed relatively small effects compared with studies using multiple-choice or fill-in-the-blank format, but still suggest that collaborative exams can aid in content retention.**

## INTRODUCTION

Activities that promote collaboration and peer interaction are linked to significant student learning (e.g., Smith *et al.*, 2009; Menekse *et al.*, 2013). One such activity is a two-stage collaborative exam, whereby students are given an opportunity to improve their understanding of a topic by first taking a test alone and then taking the test, or a portion of the test, again while interacting with a peer group. While the two-stage collaborative exam format is commonly reported to have an immediate positive effect on student test grades during the exam itself (Rao *et al.*, 2002; Cortright *et al.*, 2003; Lusk and Conklin, 2003; Giuliodori *et al.*, 2008; Leight *et al.*, 2012; Gilley and Clarkston, 2014; Mahoney and Harris-Reeves, 2017), whether it can benefit longer-term retention of course content is unclear. Some indicate that collaborative exams can improve retention of course content (Cortright *et al.*, 2003; Gilley and Clarkston, 2014; Ives, 2014), while others suggest no benefit to retention (Lusk and Conklin, 2003; Vojdanoska *et al.*, 2010; Leight *et al.*, 2012).

Some of the variability in the literature can be attributed to differences in curricula and the pedagogical approaches of the courses that have been studied. Course levels range from the introductory level (Vojdanoska *et al.*, 2010; Leight *et al.*, 2012; Gilley and Clarkston, 2014; Ives, 2014) to upper-division courses (Cortright *et al.*, 2003; Woody *et al.*, 2008), as well as graduate and/or professional school (Lusk and Conklin, 2003). The course disciplines range from science, technology, engineering, and mathematics (Cortright *et al.*, 2003; Leight *et al.*, 2012; Gilley and Clarkston, 2014; Ives, 2014) to psychology (Woody *et al.*, 2008; Vojdanoska *et al.*, 2010) to medical

**Peggy Brickman,** *Monitoring Editor*

Submitted Jul 21, 2017; Revised Jan 24, 2019; Accepted Jan 29, 2019

CBE Life Sci Educ June 1, 2019 18:ar12

DOI:10.1187/cbe.17-07-0137

[†]These authors contributed equally to the manuscript.

Present addresses: [§]Division of Biological Sciences, University of California, San Diego, La Jolla, CA 92093-0355; [¶]Department of Biology, Saint Mary's University, Halifax, NS B3H 3C3, Canada; [#]Botany Department, University of British Columbia, Vancouver, BC V6T 1Z4, Canada.

*Address correspondence to: James E. Cooke (j2cooke@ucsd.edu).

fields (Lusk and Conklin, 2003). Differences also exist in the ways groups are formed for the collaborative portion of the exam; some studies have students self-select (e.g., Woody *et al.*, 2008; Ives, 2014), while others randomly assign students to groups (e.g., Lusk and Conklin, 2003; Leight *et al.*, 2012). Indeed, the literature suggests that working with new group members (vs. returning group members) is either helpful (Gorman and Cooke, 2011) or detrimental (Liang *et al.*, 1995; Wheelan *et al.*, 2003; Opatrny *et al.*, 2014), depending on a variety of factors.

In addition to pedagogical and curricular differences, the literature on collaborative exams and retention is variable with respect to experimental design. First, the length of time used to measure retention has involved a range of time periods, from 3 days (Gilley and Clarkston, 2014) to 4 weeks (Cortright *et al.*, 2003) and up to 7 weeks (Ives, 2014). Second, some studies include an individual retest in addition to the collaborative exam to control for the test effect (e.g., Vojdanoska *et al.*, 2010; Gilley and Clarkston, 2014; Ives, 2014), while other studies do not (Cortright *et al.*, 2003; Lusk and Conklin, 2003; Leight *et al.*, 2012). The test effect is a phenomenon in which repeated exposure to content by testing can lead to increased retention (Roediger and Karpicke, 2006). Third, the retention posttests sometimes take the form of a subsequent scheduled exam (Cortright *et al.*, 2003; Lusk and Conklin, 2003; Leight *et al.*, 2012), while in other studies, the posttests are delivered without warning in an attempt to capture student knowledge "in the moment" (Vojdanoska *et al.*, 2010; Gilley and Clarkston, 2014). In light of this variation in experimental design, whereby no two studies are similar across all characteristics, it is difficult to make strong conclusions about the effect of collaborative exams on retention.

An additional aspect of study design involves test question type. Studies evaluating whether collaborative exams can improve retention of course content have used only multiple-choice questions (Lusk and Conklin, 2003; Woody *et al.*, 2008; Leight *et al.*, 2012; Gilley and Clarkston, 2014) or a mix of multiple-choice questions and some other format such as fill in the blanks (Rao *et al.*, 2002; Cortright *et al.*, 2003) or short answer (Jang *et al.*, 2017). In these studies, students were not required to generate their own answers or explanations, but rather selected the correct answer (Cortright *et al.*, 2003; Lusk and Conklin 2003; Leight *et al.* 2012; Gilley and Clarkston, 2014) or recalled a word or phrase (Vojdanoska *et al.*, 2010). However, retention may increase even further when students are required to generate their own responses to exam questions (McDaniel *et al.*, 2007). Indeed, students who answer free-response questions are able to recall significantly more information at later time points than those tested on the same material using multiple-choice format (Kang *et al.*, 2007). When students are required to generate answers on their own—as opposed to recognizing the correct answer in a field of incorrect (or less appropriate) answers—this leads to either better consolidation of material or easier retrieval of the information.

We undertook a study designed to build on previous work in an effort to elucidate whether collaborative exams can improve retention of course content. As a first, we used open-ended questions, in an attempt to allow students to generate their own responses during the exam. We used an experimental design that reduces the test effect and between-subjects variability (similar to Gilley and Clarkston, 2014; Ives, 2014). We also measured retention at two different time points (9 and 23 days) to assess retention across time, at times similar to those of Vojdanoska *et al.* (2010; 1 week), and of Leight *et al.* (2012; 3 weeks). To understand which students might be benefiting from collaborative exams, we assessed retention for students of different performance levels (similar to Leight *et al.*, 2012; Gilley and Clarkston, 2014).

## METHODS

### Course Context

Our study was conducted in an introductory biology class (Genetics, Ecology and Evolution) at the University of British Columbia during the 6-week summer semester. This course is a prerequisite for biology majors, although there were nonmajors among the 158 students enrolled in the course. Two in-class midterm exams (hereafter referred to as "midterm 1" and "midterm 2") were administered during the 6-week course and structured as two-stage collaborative exams. Midterm 1 occurred after the second week, and midterm 2 was given after the fourth week.

Students self-selected into groups of three to six students at the start of the course. Group work began on the first day of class and was used regularly every day throughout the term for clicker questions and worksheet activities. Students were informed and reminded of the benefits of maintaining the same groups for in-class discussions and group exams, but they were not forced to remain in the same groups all semester. There were 40 student groups for midterm 1 and 44 for midterm 2. Seventeen groups had a change of at least one member between midterm 1 and midterm 2; the remaining groups were consistent for midterms 1 and 2.

### Experimental Design

For this experiment, all students participated in both the control (individual retest) and treatment (collaborative retest) conditions via a quasi-experimental crossover design, administered during midterm 1 and midterm 2 (Figure 1). Each midterm consisted of three parts that occurred in the following sequence: an individual test, an individual retest, and a group retest. For analysis, we did not adjust the raw grades that the students achieved on the questions on the three tests. The individual test lasted 40 minutes and contained five questions that required short written answers; students handed in the individual test once completed, before administration of the retests.

### Retests

The individual test was followed by the individual retest (control condition). Each student had 10 minutes to individually answer one of the questions from the individual test (see the Supplemental Material). The two topics for midterm 1 were meiosis and pedigrees; the topics for midterm 2 were Hardy-Weinberg equilibrium and phylogenetic trees. Using midterm 1 as an example, during the individual retest, half of the class answered a question about meiosis and the other half answered a question about pedigrees. The inclusion of the individual retest ensured that students were tested twice in each of the control and treatment conditions, thereby eliminating the influence of a testing effect. Students then

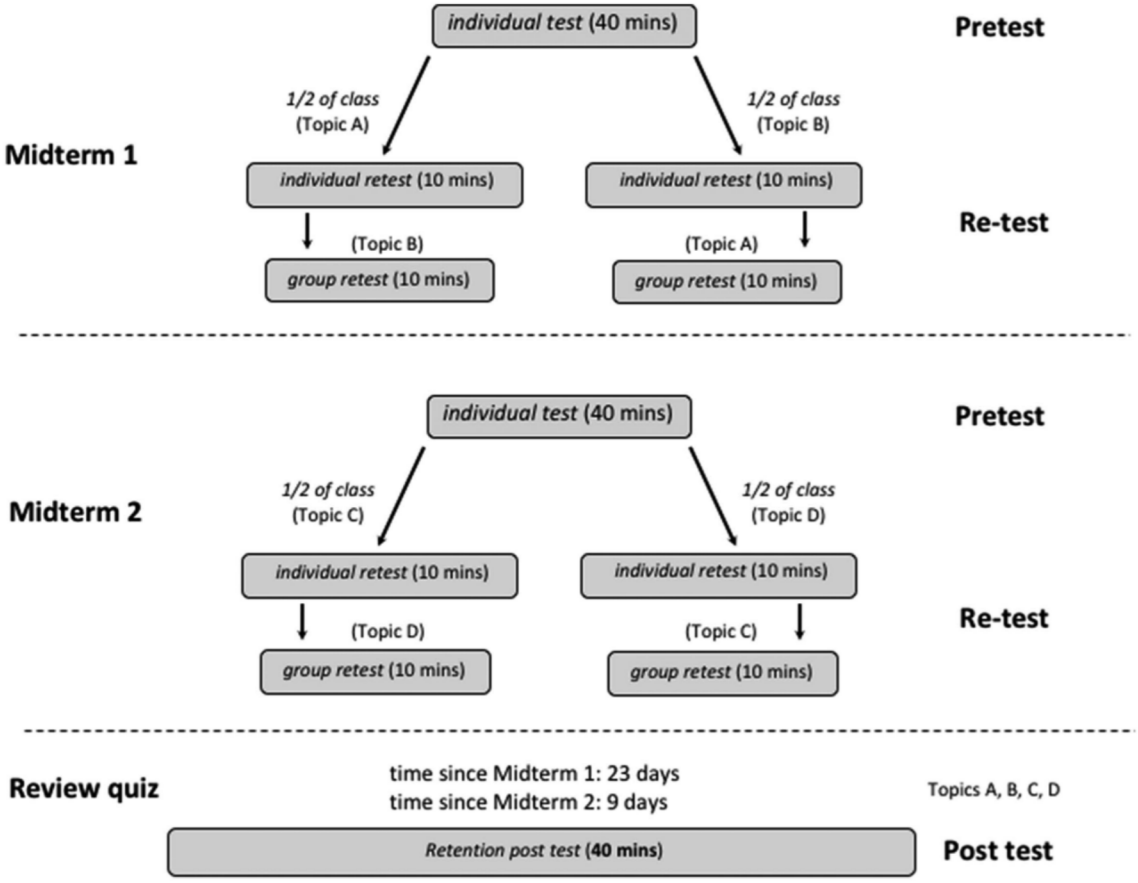


FIGURE 1. Experimental design of the study. Student's experiences of the study are listed on the left (i.e., midterm 1, midterm 2, review quiz). Elements as they pertain to the study are listed on the right (i.e., pretest, retest, posttest).

handed in their individual retest sheets and were prompted to assemble into their groups.

For the group retest (treatment condition), each group of three to six students had 10 minutes to collectively answer one question from the individual pretest. For midterm 1, this retest was isomorphic; for midterm 2, the retest was identical (more on this in the *Discussion* section). Using midterm 1 as an example, students who answered a meiosis question on the individual retest received a pedigree question in their groups, and vice versa; thus, all members in a given group had the same topic (e.g., meiosis) in their individual retests and then saw the alternate topic (e.g., pedigrees) in their group retests. In this experimental design, each student experienced both the control condition (questions rewritten individually) and treatment condition (questions rewritten in a group).

### Posttest

The retention test was administered to the students on the second-to-last class of the term (23 days following midterm 1, 9 days following midterm 2) and was presented as a final exam "review quiz" that was not worth any marks toward the final course grade. Students were not aware of the occurrence of the posttest before its administration. The retention test instructions were, "This quiz is designed to test whether you are prepared for some of the more difficult questions that you are likely to encounter on the final. This quiz does not count towards your grade. It is a tool that you can use to identify areas of strength and weakness while preparing for the final exam (next Tuesday!). I will take up the answers to this quiz after we have completed it." Students were also verbally instructed to put forth their best effort on the review quiz, because it would make the feedback session that followed much more useful. The retention test (review quiz) was written individually and consisted of isomorphic questions (see the Supplemental Material) associated with each of the four topics tested experimentally during midterms 1 and 2, plus one additional question from material presented during the last 2 weeks of class.

All parts of each midterm and the retention test were graded by the same individual. The grader was unaware of the experimental condition (control vs. treatment), as the student name(s) on the midterms were blanked out and individual and group retests were randomly mixed together during the grading. To establish a reliable grading rubric, the grader and the course instructor marked together for the first few hours, before the grader marked the rest of the midterms alone. Because there was only one person performing all of the grading, and there was no way to know the condition (treatment or control) or performance ranking (low-, mid-, or high-performing) of each exam, we can assume that there was no bias in grading.

### Data Analysis

The scores students obtained were represented as a proportion of total possible marks obtained on each question, and a binomial distribution was used in our analysis. Because we had repeated measures for individual students, a mixed modeling approach was used to avoid pseudoreplication and overinflation of the sample size. To this end, we used mixed-effects models for which student ID was included as a random, repeated factor; retest type (group or individual), midterm (midterm 1 or 2), and time (individual test and retention test) were included as fixed effects. Random effects can influence the distribution of the data, and thus the variance therein was not included in the analysis of the fixed effects of interest. Using student ID as a random effect removed any influence of inherent differences among students on the effects of test type and midterm time on student performance that were not directly associated with the tests. Significance was determined using the $\chi^2$ distribution rather than the $F$ distribution due to the binomial nature of the data. First, the entire data set was analyzed to examine the overall effects of group collaboration on individual performance by comparing individual test scores with retention test scores. The effect of collaborative exams on low-, mid-, and high-performing students was examined using the same generalized linear modeling approach as earlier, with retest type and performance category as fixed effects.

For each midterm, students were assigned to a performance category using the topic questions they completed during the individual test; these categories were created separately for each topic in both midterm 1 and midterm 2. Students were grouped into three quantiles, or tertiles, as "low," "middle," or "high" based on their performance on topic questions during the individual test. These tertiles were based on the data for each question for the whole class; the lowest 33% of the class was in the low tertile, while the top 33% of the class was in the high tertile.

Because students self-selected into groups for the group retest, there was some variation in group composition, with some groups composed entirely of students who were all in one of the performance categories. This occurred for seven groups in midterm 1 (out of 40 total groups) and 13 groups for midterm 2 (out of 44 total groups). Homogeneous groups composed of only low-, middle-, or high-performing students occurred in both midterms, with homogeneous middle or high being more common; only three groups were composed of only low-performing students.

It is important to note that this model ranks individual students based solely on their performance on different topics, not according to other metrics such as their incoming grade point average or overall grade for all questions on the individual test midterm. This approach was taken due to the highly varied topics within the introductory course; students may perform very well on some topic concepts and poorly on others. Because of the experimental design, students may be in two different performance categories for the same midterm. Sample sizes for the distribution of students across categories for each midterm and retest type are shown in Table 1. Only students who completed both midterms and the posttest were included in our analyses ($n = 125$ students out of 158 enrolled in the course). Statistical analyses were performed using R v. 3.4.4.

**TABLE 1. Number of students in each of the low, middle, and high categories for group and individual questions**

| Midterm | Retest type | Performance category | *n* |
|---|---|---|---|
| 1 | Group | Low | 47 |
| 1 | Group | Middle | 37 |
| 1 | Group | High | 41 |
| 1 | Individual | Low | 51 |
| 1 | Individual | Middle | 39 |
| 1 | Individual | High | 35 |
| 2 | Group | Low | 33 |
| 2 | Group | Middle | 38 |
| 2 | Group | High | 54 |
| 2 | Individual | Low | 27 |
| 2 | Individual | Middle | 42 |
| 2 | Individual | High | 56 |

### Student Satisfaction

To determine students' perceptions of their experience with the multistage collaborative exam, we presented them with a Likert-scale question using iClickers during the first lecture after midterm 1. The statement presented to the students was "I enjoyed the 2-stage group midterm we just wrote," with the answer options strongly agree, agree, neutral, disagree, and strongly disagree.

## RESULTS

### Effects of Group Collaboration on Retention

To examine the influence of experimental treatment on changes in student retention, we constructed a model that included an interactive effect of test timing (individual test to retention test), experimental group (control and treatment), and midterm (midterms 1 and 2) on student scores. The best model yielded two significant two-way interactions: one between midterm and experimental group and one between midterm and test timing (Table 2 and Figure 2). To explore the nature of these interactions, we examined student performance first by comparing the individual pretest scores between experimental groups and midterms, and then did a similar comparison for the retention posttest scores. First, performance on the individual test questions did not differ between midterms or experimental groups, nor was there a significant interaction between the two (midterm: $\chi^2_{1,121} = 0.36$, $p = 0.55$; treatment: $\chi^2_{1,121} = 0.14$, $p = 0.70$;

**TABLE 2. Analysis of deviance summary for the effects of retest type, time of testing, and exam on overall student performance ($n = 125$ students)**

| Fixed effect | Numerator *df* | $\chi^2$ value | *p* value |
|---|---|---|---|
| Retest type (group vs. individual) | 1 | 0.69 | 0.40 |
| Time (individual test vs. individual retention test) | 1 | 296.61 | <0.001 |
| Midterm (midterm 1 or midterm 2) | 1 | 29.89 | <0.001 |
| Retest type × time | 1 | 0.06 | 0.80 |
| Retest type × midterm | 1 | 6.37 | 0.01 |
| Midterm × time | 1 | 18.84 | <0.001 |
| Retest type × midterm × time | 1 | 0.024 | 0.88 |

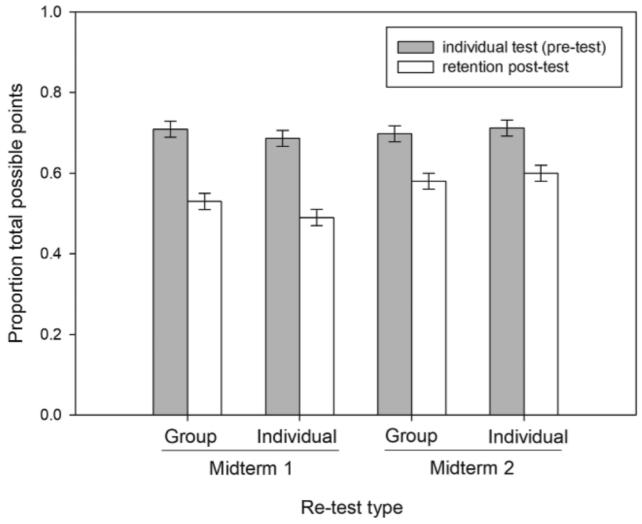


**FIGURE 2. Individual and retention test scores separated by retest type (group or individual) and by midterm (midterms 1 and 2), without distinguishing between performance type. Individual and retests occurred 23 and 9 days before the retention posttest for midterms 1 and 2, respectively. Bar heights represent means, and error bars are SE.**

interaction: $\chi^2_{1,121} = 2.67$, $p = 0.10$; Figure 2). However, there was a significant interaction between midterm and experimental group on scores on retention posttest question ($\chi^2_{1,121} = 3.89$, $p = 0.048$; Figure 2). This interaction is attributable to the fact that, for midterm 1, which was 23 days before the posttest, students scored better on retention posttest topics that they rewrote in groups compared with topics rewritten individually ($\chi^2_{1,123} = 4.37$, $p = 0.037$; Figure 2). This was not the case for midterm 2, which was 9 days before the posttest, for which students performed equally well on retention posttest topics that they had rewritten in groups or individually ($\chi^2_{1,123} = 0.71$, $p = 0.40$; Figure 2). Performance on the retention posttest was higher for material associated with midterm 2 compared with midterm 1 for both the group ($\chi^2_{1,123} = 12.93$, $p < 0.001$; Figure 2) and individual ($\chi^2_{1,123} = 44.16$, $p < 0.001$; Figure 2) conditions.

### Student Performance Categories

To further investigate the effects of collaborative exams on content retention, we compared retention posttest scores across the high, middle, and low student performance categories. For midterm 1, there was a weak, nonsignificant interaction between performance category and retest type (Table 3 and Figure 3a), driven by slightly better performance of students ranked as low on the topics associated with the group retest compared with the individual retest (Tukey post hoc test: odds ratio [OR] = 1.49, $p = 0.086$; Figure 3a). Scores did not differ between retest types for students in the middle or high categories on the first midterm (Table 3 and Figure 3a). By contrast, for midterm 2, there was a significant interaction between performance category and retest type on retention posttest scores (Table 3 and Figure 3b). This interaction included slightly higher scores for mid-performing students on topics for which students did their retest in groups (Tukey post hoc test: OR = 1.51, $p = 0.059$; Figure 3b). Scores did not differ between retest types for students in the low or high categories on the second midterm (Figure 3b).

**TABLE 3. Analysis of deviance summary for the effects of retest type, time of testing, and exam on overall student performance ($n = 125$ students)**

| Fixed effect | Numerator *df* | $\chi^2$ value | *p* value |
|---|---|---|---|
| Midterm 1 | | | |
| Tertile (low, middle, high) | 2 | 35.45 | <0.001 |
| Retest type (group vs. individual) | 1 | 2.47 | 0.11 |
| Retest type × tertile | 2 | 5.08 | 0.079 |
| Midterm 2 | | | |
| Tertile (low, middle, high) | 2 | 111.76 | <0.001 |
| Retest type (group vs. individual) | 1 | 0.024 | 0.11 |
| Retest type × tertile | 2 | 11.00 | 0.004 |

### Student Satisfaction

During the first lecture following the multistage collaborative exam for midterm 1, we administered a Likert-scale survey to the students using iClickers. In response to the statement

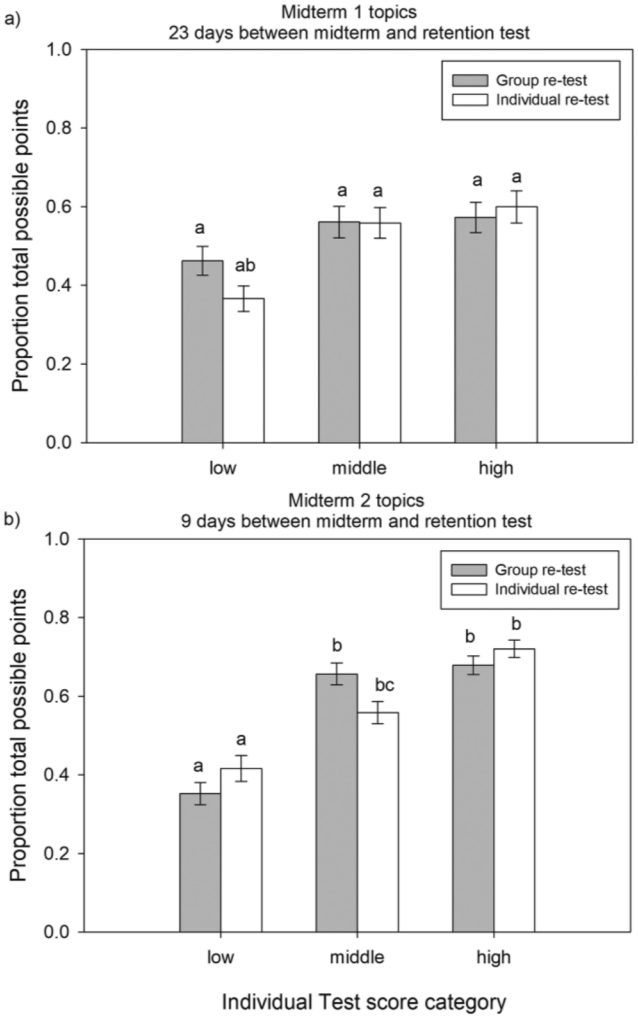


**FIGURE 3. Retention test scores for low-, middle-, and high-performing students for (a) midterm 1 and (b) midterm 2. Bars represent means, and error bars are SE. Letters above the bars indicate groupings in Tukey post hoc tests; different letters indicate significant differences in pairwise tests of least-squares means.**

"I enjoyed the 2-stage group midterm we just wrote," 96 out of 123 (78%) participating students reported generally positive perspectives, 13 (11%) reported neutral feelings, and 14 (13%) reported generally negative feelings about the experience. As there were 156 students who took midterm 1, the 123 students participating in the survey accounted for 78% of those surveyed.

## DISCUSSION

Our study suggests that collaborative testing improved retention of course content at 23 days after the midterm for the class as a whole, although there were no statistically significant differences between group and individual scores for low-, mid-, or high-performing students. At 9 days after the midterm, there was no significant effect of collaborative testing on retention for the class as a whole or for low- and high-performing students specifically, but there was a significant improvement in retention for mid-performing students.

### Collaborative Exams and Retention

A generalizable timeline for the impact of collaborative testing on retention based on the literature remains elusive, because the results have been mixed for the relatively few studies that have measured retention in the days and weeks following a collaborative test. In the short term, Gilley and Clarkston (2014) measured retention at 3 days and found significant improvement for students across all performance categories when tested collaboratively; Ives (2014) found a similar effect of collaborative testing measured at 1–2 weeks. Meanwhile, Vojdanoska *et al.* (2010) measured retention at 7 days and found no benefit to group testing. In the longer term, Cortright *et al.* (2003) observed retention benefits of collaborative exams measured at 4 weeks, while others found no benefit at 3 weeks (Lusk and Conklin, 2003; Woody *et al.*, 2008) and 6–7 weeks (Ives, 2014). In the present study, the observation that collaborative testing improved retention of course content for the class as a whole at 23 days aligns with the findings and timeline of Cortright *et al.* (2003). Ives (2014) is one of the few to measure retention twice (at 1–2 and 6–7 weeks) and, interestingly, he found that collaborative testing improved retention at 1–2 weeks, but not at 6–7 weeks, while the current study found improved retention at 23 days, but not at 9 days (for the class as a whole). However, Ives's second time point, 6–7 weeks, is roughly twice the length of our "longer-term" time point, making it difficult to directly compare the effects of different timelines for the two studies.

At the outset of this study, we expected that any observed improvement in retention in the collaborative testing treatment would follow a general pattern of being highest when measured at the time point closest to the midterm and decreasing as more time elapsed. This would be consistent with collaborative testing studies with similar experimental designs (Ives, 2014), as well as literature on the decay of memory performance over time (Sayre and Heckler, 2009) and the consistent rates of decreased retention over time independent of how "active" or "passive" the classroom environment was deemed to be (Deslauriers and Wieman, 2011). However, our results mostly did not meet our predictions: for the class as a whole, students scored better on retention posttest topics that they rewrote in groups compared with topics rewritten individually for midterm 1 (23 days to posttest), but this effect was not seen in midterm 2 (9 days to posttest), although posttest grades were generally higher for material associated with midterm 2. Furthermore, the benefit observed for mid-performing students at 9 days was lost by 23 days—consistent with our expectations of decreased retention over time. Perhaps most interesting, the benefit observed for low-performing students at 23 days was not present at 9 days. We would have expected the retention benefit to be present starting from the day of the treatment (the collaborative exam) and to persist up to 23 days.

It is also plausible that the differences we observed in retention timing at 9 and 23 days are not entirely due to the time since the exam, but also to differences in the material that was presented at each exam (i.e., genetics for midterm 1 and evolution for midterm 2). To our knowledge, this result (though weak and nonsignificant) is novel in the collaborative testing literature. We believe that this area should be the subject of future study with a more direct design: the same topic measured at different points in time.

### Differential Benefits for Low-, Mid-, and High-Performing Students

While the majority of studies have evaluated retention for the entire student population together, few studies have examined whether there were differential retention benefits for students of different performance levels. While Mahoney and Harris-Reeves (2017) did not evaluate retention, they found grades improved for low- and mid-performing students on a collaborative test (compared with identical questions answered previously as individuals), but not high-performing students—the authors speculated this might be due to high performers bowing to peer pressure to choose an incorrect answer when in a group setting. The authors also found that collaborative exams improved performance on questions that required higher-order cognitive skills for students of all performance levels. In our study, we found that both low- and mid-performing students benefited from the collaborative exam, although this differed depending on the time between the exam and the retest. The results for the low-performing students were not statistically significant, but there was a trend toward higher performance for students rewriting the test in groups on the first midterm, and a moderate effect size associated with treatment of 0.18 (Table 4). In addition, the mid-performing students showed larger gains for midterm 2, whereby the effect size of the treatment on retention tests was relatively large at 0.42 (Table 4). These findings differ from those of Gilley and Clarkston (2014), who found that collaborative exams improved retention for all categories of students, as well as those of Leight *et al.* (2012), who found that collaborative exams offered no retention benefits for students at any letter grade.

Our best explanation for our results regarding different performance categories is 1) that students were independently assigned to performance categories for each midterm and 2) that the content topics for each midterm were different. With respect to performance categories, the actual students in each performance category for a given midterm is different; being classified as mid-performing based on the collaborative question for midterm 1 had no bearing on a given student's classification for midterm 2. While some of the students were

TABLE 4. Effect sizes for the influence of retest type on individual retention test scores

| Student performance category | Effect size[a] |
|---|---|
| Midterm 1 | |
| Overall | 0.1 |
| Low | 0.18 |
| Middle | 0.017 |
| High | 0.010 |
| Midterm 2 | |
| Overall | −0.05 |
| Low | −0.22 |
| Middle | 0.42 |
| High | −0.21 |

[a]Positive values indicate higher scores on questions associated with group retests.

classed as mid-performers in both midterms, there are some students new to this designation for midterm 2. With respect to content, midterm 1 contained only questions about genetics, while midterm 2 contained questions about evolution. The questions used for the retests for both midterms were carefully chosen to be as close in difficulty and reliability as possible based on past student performance in other iterations of the course. However, it is possible that students categorized as mid-performers for midterm 2 had a more difficult time with the topic of evolution than did the mid-performing students for midterm 1 with the topic of genetics. These explanations are also valid for low-performing students having retention benefits at 23 days but not 9 days: different students and different subject matter.

In addition, to our knowledge, the current study is unique in using only open-ended, short-answer questions for all parts of the study (pretest, retest, and posttest). All other studies evaluating whether collaborative exams can improve retention of course content have used only multiple-choice questions (Lusk and Conklin, 2003; Leight *et al.*, 2012; Gilley and Clarkston, 2014) or a mix of multiple-choice questions and some other format such as fill-in-the-blank (Rao *et al.*, 2002; Cortright *et al.*, 2003), short-answer (Jang *et al.*, 2017), or essay questions (Woody *et al.*, 2008). To our knowledge, no other study has used open-ended, short-answer questions on the posttest portions of their studies. Using only short-answer, written-response questions requires students to generate their own responses for all individual and collaborative stages of the test. When students have to generate their own responses, subsequent retention of information is greater than when students have to recognize a correct response, as is the case with multiple-choice questions (Kang *et al.*, 2007; McDaniel *et al.*, 2007). It is possible that the combination of answering short-answer questions for different topics at different time scales impacted retention for different student performance categories in ways that were unexpected and not explicitly examined in this study.

While our posttest questions were all isomorphic to the individual pretest questions, the retest questions for midterm 2 were either identical (Hardy-Weinberg equilibrium question) or nearly identical (phylogenetic tree; see the Supplemental Figures). This happened as a result of an error in preparing for the second midterm exam. It is possible that this difference would make it more likely that students would remember the answers to specific questions (e.g., whether pigs or whales are more closely related to camels on a phylogenetic tree). However, because posttest questions were all isomorphic, we cannot see any benefit (or detriment) that committing specific details to memory would have for answering a question with different details (e.g., a totally different phylogenetic tree). As such, we feel that this difference is minor and doubt that it contributes to our observed effects.

### Student Perspectives of Collaborative Exams

In addition to the learning benefits that can be conferred by collaborative exams, students in our study generally reported enjoying the experience. Our data on student perspectives of collaborative exams—in which 78% of students reported positive generally positive experiences—are very similar to the data obtained by Rieger and Heiner (2014) in an introductory physics class. Rieger and Heiner (2014) found that students were generally positive in their collaborative exam experience 76% of the time, neutral 14% of the time, and had negative experiences 10% of the time. Similarly, many studies over the past 15 years have reported positive experiences during collaborative exams, from reductions in test anxiety to perceptions of enhanced learning (for a review, please see LoGiudice *et al.*, 2015). Taken together, the data strongly suggest that collaborative exams offer benefits beyond grade improvement and retention of course content.

### Summary and Future Directions

Our study indicates that two-stage collaborative exams can improve retention of course content up to 23 days after the exam and, when looking at specific student performance categories, that both low-performing and mid-performing students benefit from collaborative exams, depending on the timing of the posttest. However, the lack of consistent methodologies within the collaborative testing literature makes it difficult to directly compare the present study with others and impossible to broadly summarize the impact of collaborative testing on retention. That some studies found improved retention when testing collaboratively, while others have not, could be due to a number of factors that vary from study to study and are, in some cases, incompletely reported: student group dynamics, methods of administering posttests, timing of posttest administration, ranking of student performance, and type of questions asked (multiple-choice vs. open-ended questions; identical vs. isomorphic). To facilitate comparisons with existing studies, future collaborative testing studies should 1) adopt a robust experimental design that accounts for the testing effect and time on task (e.g., this study; Gilley and Clarkston, 2014; Ives, 2014); and 2) thoughtfully consider and clearly report factors such as how student groups are created and managed, how topics and questions are chosen, and how student performance groups are created. In addition to these considerations, future studies aimed specifically at measuring retention should include multiple time points (to better capture patterns of retention for different groups and over time) and choose time points similar to existing studies (e.g., 3 days, 1 week, 4 weeks, 7 weeks) to facilitate comparison with existing literature and contribute to our understanding of how students learn.

## ACKNOWLEDGMENTS

We thank Dr. L. McDonnell for helpful comments on an earlier draft of the article.

## REFERENCES

Cortright, R. N., Collins, H. L., Rodenbaugh, D. W., & DiCarlo, S. E. (2003). Student retention of course content is improved by collaborative-group testing. *Advances in Physiology Education*, *27*(1–4), 102–108.

Deslauriers, L, & Wieman, C (2011). Learning and retention of quantum concepts with different teaching methods. *Physical Review Special Topics Physics Education Research*, *7*(1). 010101-1–010101-6.

Gilley, B. H., & Clarkston, B. (2014). Collaborative testing: Evidence of learning in a controlled in-class study of undergraduate students. *Journal of College Science Teaching*, *42*(3), 83–91.

Giuliodori, M. J., Lujan, H. L., & DiCarlo, S. E. (2008). Collaborative group testing benefits high- and low-performing students. *Advances in Physiology Education*, *32*(4), 274–278.

Gorman, J. C., & Cooke, N. (2011). Changes in team cognition after a retention interval: The benefits of mixing it up. *Journal of Experimental Psychology: Applied*, *17*(4), 303–319.

Ives, J. (2014). Measuring the learning from two-stage collaborative group exams. Retrieved November 11, 2017, from arXiv:1407.6442.

Jang, H., Lasry, N., Miller, K., & Mazur, E. (2017). Collaborative exams: Cheating? Or learning? *American Journal of Physics Education*, *85*(3), 223–227.

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4), 528–558.

Leight, H., Saunders, C., Calkins, R., & Withers, M. (2012). Collaborative testing improves performance but not content retention in a large-enrollment introductory biology class. *CBE—Life Sciences Education*, *11*(4), 392–401.

Liang, D. W., Moreland, R., & Argote, L. (1995). Group versus individual training and group performance: The mediating role of transactive memory. *Personality and Social Psychology Bulletin*, *21*(4), 384–393.

LoGiudice, A. B., Pachai, A. A., & Kim, J. A. (2015). Testing together: When do students learn more through collaborative tests? *Scholarship of Teaching and Learning in Psychology*, *1*(4), 377–389.

Lusk, M., & Conklin, L. (2003). Collaborative testing to promote learning. *Journal of Nursing Education*, *42*(3), 121–124.

Mahoney, J. W., & Harris-Reeves, B. (2017). The effects of collaborative testing on higher order thinking: Do the bright get brighter? *Active Learning in Higher Education*, *20*(1), 25–37. doi:10.1177/1469787417723243

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morisette, N. (2007). Testing the test effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4–5), 494–513.

Menekse, M., Stump, G. S., Krause, S., & Chi, M. T. H. (2013). Differentiated overt learning activities for effective instruction in engineering classrooms. *Journal of Engineering Education*, *102*(3), 346–374.

Opatrny, C., McCord, M., & Michealsen, L. (2014). Can transferable team skills be taught? A longitudinal study. *Academy of Educational Leadership Journal*, *18*(2), 61–72.

Rao, S. P., Collins, H. L., & DiCarlo, S. E. (2002). Collaborative testing enhances student learning. *Advances in Physiology Education*, *26*(1–4), 37–41.

Rieger, G. W., & Heiner, C. E. (2014). Examinations that support collaborative learning: The students' perspective. *Journal of College Science Teaching*, *43*(4), 41–47.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychology Science*, *17*(3), 249–255.

Sayre, E. C., & Heckler, A. F. (2009). Peaks and decays of student knowledge in an introductory E&M course. *Physical Review Special Topics Physics Education Research*, *5*(1). 013101.

Smith, M. K., Wood, W. B., Adams, W. K., Wieman, C. E., , Knight, J. K., Guild, N., & Su, T. T. (2009). Why peer discussion improves student performance on in-class concept questions. *Science*, *323*(5910), 122–124.

Vojdanoska, M., Cranney, J., & Newell, B. R. (2010). The testing effect: The role of feedback and collaboration in a tertiary classroom setting. *Applied Cognitive Psychology*, *24*, 1183–1195.

Wheelan, S. A., Davidson, B., & Tilin, F. (2003). Group development across time: Reality or illusion? *Small Group Research*, *34*(2), 223–245.

Woody, W. D., Woody, L. K., & Bromley, S. (2008). Anticipated group versus individual examinations: A classroom comparison. *Teaching Psychology*, *35*(1), 13–17.