What Is The Impact Of Collaborative Exams On Learning And Attitudes In Introductory Astronomy Classes?

Scott T. Miller, Sam Houston State University, USA
C. Renée James, Sam Houston State University, USA

ABSTRACT

We present results of a two-semester study to gauge the impact of collaborative two-stage exams on student learning and attitudes in university-level introductory astronomy classes for non-science majors. In the collaborative two-stage exam setting, students first completed an exam individually, and then they reconsidered a subset of exam questions within their previously established groups, discussing the questions with their peers to arrive at a common answer.Students took three to four exams during the semester using this format. At mid-semester, we surveyed the students to gauge their attitudes about collaborative work and its perceived influence on their exam preparation and performance. At the end of the semester, students sat an individual-only final exam, which contained all previous collaborative-phase questions, as well as a subset of questions seen only on the individual portions of the exams. When we compare the normalized gain on final exam questions that were included in the collaborative portions to that on questions found in only the individual portions, we find higher normalized gains in general for questions encountered on the collaborative portions of the exams. These gains are accompanied by a statistically significant effect size (Cohen's d). We note, however, that this improved performance appears to be dependent upon several factors. Those factors might include diminished retention over time, the assessment of overly complex concepts, and concept saturation. Our mid-semester survey indicates that the collaborative experience appears have a positive influence on their overall attitudes and their study habits.

Keywords*:* Astronomy Education Research; College Teaching; Collaborative Learning; Attitudes

The layout of the standard classroom or lecture hall on a university campus immediately betrays its intended purpose. A professor is to stand at the lectern in the front and convey information to the students, who are to be seated in orderly rows, silently taking notes. After providing this form of instruction to their classes, however, many science professors then head to their offices or labs, where discussions with colleagues about their latest project or upcoming proposals resume. Scientists around the globe rely on collaborative efforts, and for good reason. As Beaver (2001) points out, collaborations aid in tackling bigger questions than individuals can tackle, improve efficiency and productivity, help individuals keep their own work more focused, spark enthusiasm, and provide access to knowledge and skills that each member does not personally possess.
Kenneth Bruffee (1987) summarizes the collaborative experience:

> *You know how it goes. Joe gets and idea and sketches it out in a couple of pages, Mary says, hey, wait a minute – that makes me think of… Then Fred says, but look, if we change this or add that…In the end everyone, with a little help from his or her friends, exceeds what anyone could possibly have learned or accomplished alone.*

Although professional collaborations have been practically indispensable to researchers for centuries, promoting collaborations in an undergraduate setting is a relatively new phenomenon. The term "collaborative learning" emerged only around 1970, slightly preceded by the similar concepts of "peer learning" and "peer instruction," which began appearing in the literature around 1960.

Also emerging over the past few decades is the two-stage collaborative exam, which has slowly grown in popularity in a wide range of courses. In a two-stage collaborative exam, students first take a test individually, and then they work in groups to come to a consensus on the answers to some subset of the individual exam questions (e.g., Stearns, 1996; Yuretich, Khan, Leckie & Clement, 2001).

The motivation for incorporating such assessments varies somewhat from instructor to instructor. For instance, in a research methods and statistics class, Stearns (1996) noted – unsurprisingly – that students appeared less interested in why they missed questions than in the grade itself. Consequently, the standard approach of handing back an exam and "debriefing" failed to engage students with the subject or persuade them to consider the rationale for their responses. In an introductory oceanography class for non-science majors, Yuretich et al. (2001) employed two-stage cooperative exams as part of a larger-scale push to incorporate more active-learning strategies in a large-enrollment course where students had historically considered "dry as dust." More recently, Knierem, Turner and Davis (2015) incorporated two-stage exams in introductory geology courses with an eye to improving "attendance, engagement, and student learning."

Regardless of the subject or level, one common motivation for instructors to incorporate two-stage collaborative exams is to improve student learning outcomes. Although a small study, Stearns (1996) noted the contrast in final exam performance between students who had simply been told what the answers were on previous exams and those who were required to discuss in small groups their rationale for choosing answers on previous exams. Specifically, the class that had simply been told the answers to previous exams scored an average of 10% lower on the final exam than the class that had to discuss exam answers and defend their rationales. However, Stearns cautions that this was only "quasi-experimental," with a number of potential biases.

Jang, Lasry, Miller, and Mazur (2017) studied an introductory calculus-based physics class at Harvard University and made two somewhat surprising finds. The first was that exam performance improved even for the strongest students when they collaborated with their peers compared to their individual performance, refuting the suggestion that the best students were simply providing the group answers. The second surprising result was that groups occasionally arrived at a correct answer collectively even when none of the group members did so individually.

How long does this collaborative advantage last, though? The few studies that have explored the impact of collaborations on sustained student learning at various time intervals have yielded mixed results. For instance, Cortright, Collins, Rodenbaugh, and DiCarlo (2003) found higher gains in learning via group exam questions over individual-only exam questions in an undergraduate physiology course where questions were mostly multiple choice, fill in the blanks, and short answer. In their study, four weeks elapsed between the initial exposure to a test question and its appearance on a retest, suggesting that at least over a period of a month, collaborative exams improve student retention of course concepts.

Gilley and Clarkston (2014) found similar results in their study of collaborative exams in an undergraduate science course where they even controlled for time on task. Their interval, however, between test and retest was brief – only three days – and students were not forewarned of the retest.

In contrast, for a large enrollment introductory biology class, Leight, Saunders, Calkins, and Wither (2012) found no statistical difference in gains between questions assessed as part of a collaborative exam versus those answered individually. Complicating the result was the fact that the questions in the retesting situation were not identical to those asked previously. Instead, they covered similar topics and were geared at a similar cognitive level.

Meanwhile, in an introductory, calculus-based physics course, Ives (2014) found that, when retested on a number of diagnostic questions, students performed better on questions addressed as a group compared to those answered individually. This improved performance, however, was seen only for material retested after a week or two. The performance on group and individual items, however, was similar when more than six or seven weeks elapsed between tests.

The collaborative nature of two-stage exams can impact more than simply learning outcomes, but there is scant work exploring the affective nature of collaborative exams. One notable exception is that of Meseke, Natziger, and Meseke (2010), an investigation into not just the learning gains, but also the attitudes of students immersed in a collaborative environment in a neuroanatomy course in a chiropractic college. Interestingly, they find no collaborative advantage for long-term learning or retention as compared with a control group. However, qualitative survey results reveal lower test anxiety and greater satisfaction with the course.

The wide variation in both the methods and results of the previous studies leaves several questions unresolved. Many cohorts studied have been majors in the discipline, so it is difficult to know whether their results can be extrapolated to non-majors or even to a different type of university (R1 versus, say, public regional). Furthermore, most studies did not address the impact of the collaborative exams on attitudes. The main exception focused on a highly specialized class with students of similar academic and professional aspirations. In fact, studies done to gauge the impact of collaborative exams on both learning and attitudes in science courses designed for non-majors appear to be absent.

With this in mind, we have ventured to answer the following questions: Does the two-stage collaborative exam format impact longer-term student learning in non-majors in introductory astronomy classes? If so, does the impact change over time? Does the two-stage collaborative exam impact other areas of the students' academic life, such as study habits or confidence in a group setting?

METHODS

This study was carried out over two semesters at the authors' institution, a public regional university where approximately half the students are first-generation college students and two-thirds are considered "at risk." We gathered numerical data from exam items as well as qualitative data from survey questions.

Our preliminary data were obtained during a 15-week semester of Solar System Astronomy in the Fall 2017 semester, followed by further investigation during two introductory astronomy courses (Solar System Astronomy and Stars & Galaxies) in the Spring 2018 semester. The two Fall 2017 sections of Solar System Astronomy and two Spring 2018 sections of Stars & Galaxies met twice a week (Tuesdays and Thursdays), with each class meeting lasting 75 minutes. The enrollment of each section was approximately 100 students. The two Spring 2018 sections of Solar System Astronomy met three times a week (Mondays, Wednesdays, and Fridays), with each class meeting lasting 50 minutes. The enrollment of each section was approximately 80 students.

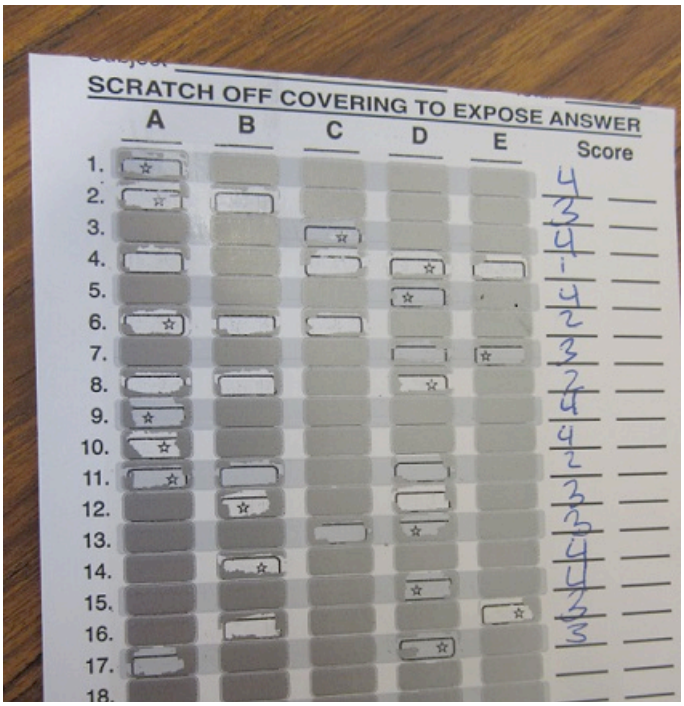
Groups were determined within the first week of each semester. Groups consisted of five to seven members sorted by classification, gender, and declared major. The Spring 2018 Solar System groups were also sorted based on performance on a pre-semester concept inventory. Each group had approximately one senior and/or junior, one to two sophomores, and two to four freshmen. In each group, a variety of majors was represented. Groups were also organized so that they were either all men, all women, or majority women. This step was taken in an effort to encourage everyone to have a voice, as it has been well documented in STEM classrooms (e.g., Finelli, Inger, and Mesa, 2011; Tonso, 2006) that women tend not to speak up as frequently in groups with a larger number of men than women. Students who had taken a previous class with the instructor were placed into separate groups, as we felt that their insight into the workings of the class was potentially beneficial.

For all classes in our study, group work was an everyday occurrence. For a few weeks before the first exam, the groups had the opportunity to work on a number of in-class activities and to discuss numerous personal response questions with their peers.

During the Fall 2017 semester, students sat for three in-class exams during weeks 5, 9, and 13 of the semester. Each of the three in-class exams consisted of two stages: an individual portion and a group portion. For the individual portion, students had 50 minutes to complete 50 multiple-choice questions individually, and this portion was worth 75% of each overall exam grade. Students who finished early were instructed to hand in their Scantron forms and their exam papers and to sit and wait until time was called. For the group portion, students gathered into their assigned groups and were given ten of the exam questions and an Immediate Feedback Assessment Technique (IF-AT) scratch-

off card (See Figure 1). Groups had 25 minutes to discuss these questions and arrive at an agreed-upon answer. Full credit was awarded for each question answered correctly on the first try, with partial credit awarded for subsequent attempts. The group portion of the exam was worth 25% of the exam grade.

Figure 1. A sample IF-AT scratch-off response card.



During the sixteenth week of the Fall 2017 semester, we administered a two-hour, individual-only final exam. The final exam consisted of 100 multiple-choice questions, including all 30 of the group questions previously encountered in exams 1, 2, and 3, along with thirty other non-group questions, ten from each of the three in-class exams. In total, 60 of the 100 items on the final exam were copied from the in-class exams.

During this semester, students had access to the previous exams and their answer keys. In fact, we encouraged students to pick up their individual exams for studying purposes, informing them that twenty questions from each in-class exam would appear on the final exam. They were not given any information about which questions would be included.

In the Fall 2017 Solar System Astronomy course, there were some exam questions that were time-dependent. For instance, a question asked in September (Exam 1) about the location of Earth in its orbit and the constellations most likely to be viewed at night had a different correct answer when asked again on December's final exam. Questions about "today's" lunar phase were also similarly affected. Although included in the final exam, we removed from the analysis any questions that had to be altered in any way between their appearance on the in-class exam version and the final exam version. Overall 27 of 30 original group questions were considered in the final analysis, along with 27 individual questions with similar student success rates.

The format of the Spring 2018 Stars and Galaxies course was similar to the Fall 2017 Solar System Astronomy course, given that they were taught by the same professor. Specifically, three in-class exams were administered during weeks 5, 8, and 12, followed by a cumulative final exam during week 16. Again, each individual exam consisted of 50 multiple-choice questions considered individually (75% of the total exam grade), with ten of those questions re-considered in the group setting (25% of the total exam grade).

Also, during the Spring 2018 semester, Solar System Astronomy students sat for four in-class exams during weeks 3, 7, 11, and 15, with a final exam administered during week 16. These exams contained 30 multiple-choice questions each, with 7 of those considered as a group. The group questions constituted 20% of each exam grade.

A key difference between the classes was that for the Fall 2017 Solar System and Spring 2018 Stars and Galaxies class, students were asked to retrieve previous exams and answer keys from the faculty member's office outside of class times. Approximately one-third of the exam papers were left in the faculty member's office at the end of each semester, but those students still had the ability to access the questions and the keys through their group members. Students in the Spring 2018 Solar System Astronomy classes, on the other hand, received their exam papers in class, and the key was provided through the university's learning management system.

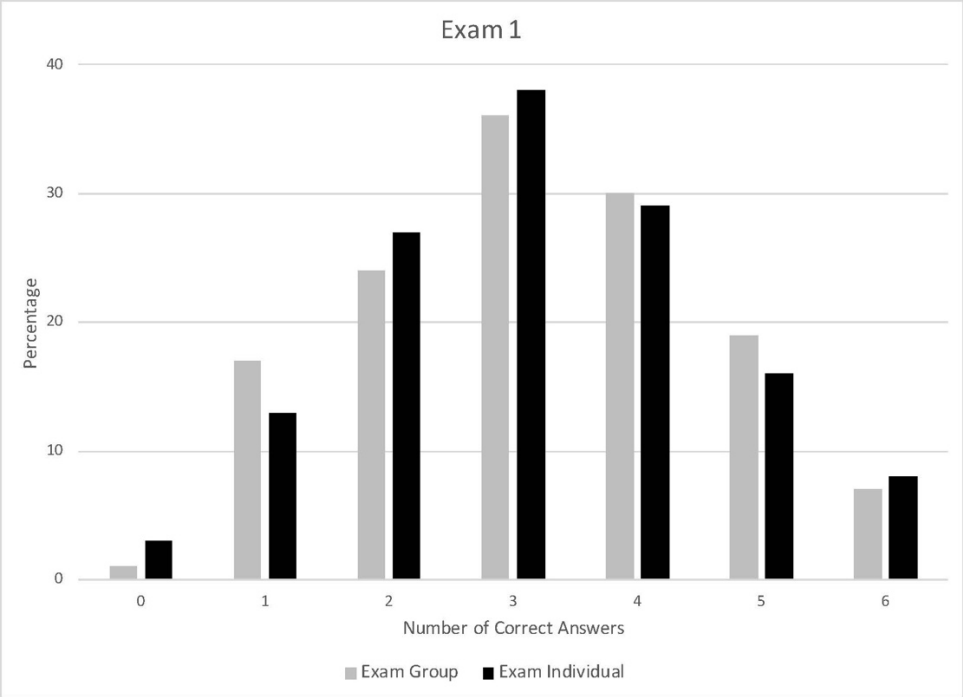
When choosing which individual-only questions to include in the final exams, we looked not only at the percentage of students who chose the correct answer for each item, but also at the distribution of correct/incorrect answers as a whole for the two types of questions. We first performed a t-test on the exam item data, comparing group results with individual results on each exam to ensure that the performance on these two question types was not statistically distinct before the final exam (Table 1). Individual questions used for comparison were chosen such that the combined averages on those questions were similar to the combined averages for the group questions. All P values were above a threshold limit of 0.05, confirming that our two sets of exam items were statistically similar.

Table 1. P values for paired sample t-tests, comparing student performance on group questions vs. individual questions included on various exams.

		n	P value
Fall 2017 Solar System	Exam 1	167	0.324
	Exam 2	168	0.176
	Exam 3	163	0.932
Spring 2018 Solar System	Exam 1	134	0.766
	Exam 2	133	0.826
	Exam 3	136	0.553
	Exam 4	139	0.884
Spring 2018 Stars & Galaxies	Exam 1	155	0.246
	Exam 2	153	0.672
	Exam 3	154	0.509

As Figure 2 illustrates, the percentage of students choosing the correct response for 0 of the 6 group questions on the individual portion of Exam 1 is similar to the percentage of students choosing 0 of the 6 non-group questions in Exam 1. Likewise, the percentage of students who chose the correct response to 1 of the 6 group questions in Exam 1 is similar to the percentage of students who chose the correct response to 1 of the 6 non-group questions in Exam 1, etc.

Figure 2. Example distribution of student success rates on the questions encountered on both the individual and group portions, compared to a comparable set of questions encountered only on the individual portion. Note the similar distribution of scores.



The final exam for Stars and Galaxies in the Spring 2018 semester contained all 30 group questions and 30 individual questions with statistically similar success rates, along with 40 other questions covering material from the entire semester. The final exam for Solar System Astronomy that semester contained 27 group questions and 27 individual questions, along with 46 additional questions.

Preliminary results from the Fall 2017 Solar System Astronomy class compelled us to gather qualitative data during the Spring 2018 semester. Approximately half way into the Spring 2018 semester, we asked students in both courses to complete an online survey to rate their experience with group work as a whole, and collaborative exams in particular. Students rated their level of agreement to several statements using a simple Likert 5-point scale. Statements alternated between positive (e.g., "During the group portion of the exams, the group discussion helps me better understand how to arrive at the answer.") and negative (e.g., "My group members often dismiss what I have to say.") in an effort to minimize the tendency to give the same answer throughout. We provided an open-ended question allowing for general comments on the collaborative experience. Overall, 136 of our 299 students from the Spring 2018 semester responded to the survey.

RESULTS AND ANALYSIS

Quantitative Results

We present the average normalized gains (as per Hake, 1998) for each question set by course and exam (See Figure 3). The average normalized gain is determined by taking the percent gain obtained by the participants and dividing it by the maximum percentage gain possible. This can be calculated as

$$\langle g \rangle = \frac{\%post - \%pre}{100\% - \%pre}$$

Where *%pre* is the average score for a question set on an exam and *%post* is the average score for a question set on the final exam. We also performed a t-test on the final exam item data (see Table 2) similar to the one performed on the previous exam data to compare the performance on the group questions with the performance on the individual questions on the final exam.

Figure 3. Normalized gains for group (light shading) and individual (dark shading) items that appeared on in-class exams indicated and on the final exam in the (a) Fall 2017 Solar System course, (b) Spring 2018 Solar System course, and (c) Spring 2018 Stars and Galaxies course.

Figure 3a

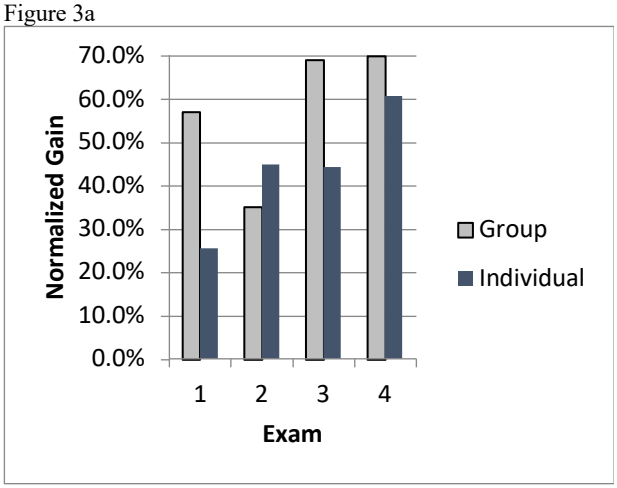


Figure 3b

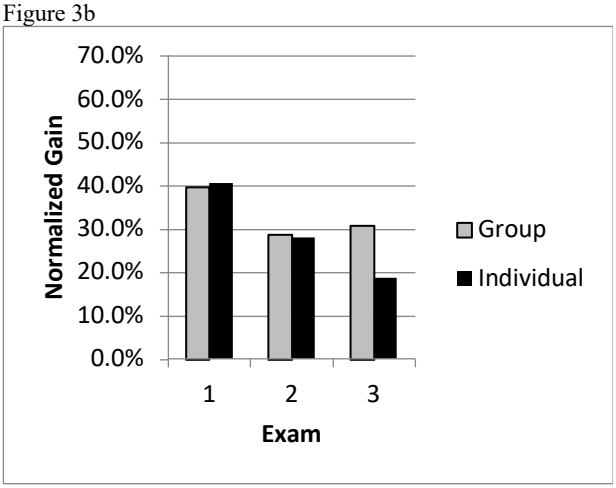


Figure 3c

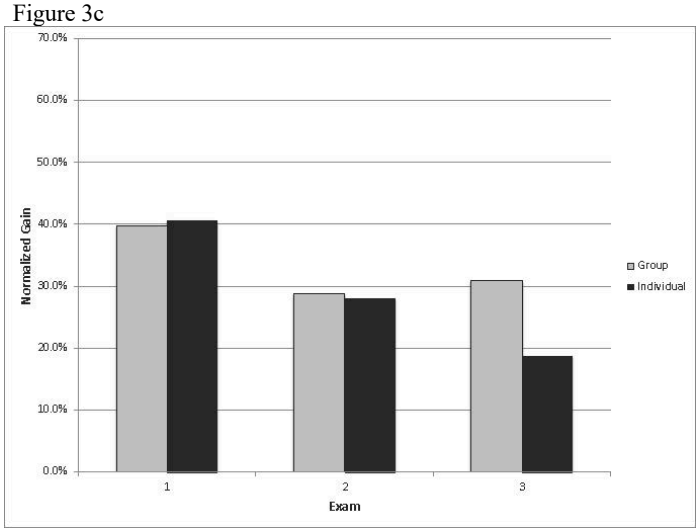


Table 2. P values for paired sample t-tests and Cohen's d effect size. The t-tests compared student performance on group questions to student performance on individual questions that were repeated on the final exam. We consider P values < 0.05 (shaded) and effect sizes of 0.20 or more significant. Significant effect sizes are shaded, and medium effect sizes are in boldface.

		N	P value	Effect Size
Fall 2017 Solar System	Exam 1	167	0.387	0.07
	Exam 2	168	<.001	0.33
	Exam 3	163	<.001	0.22
Spring 2018 Solar System	Exam 1	134	<.001	**0.63**
	Exam 2	133	0.010	0.25
	Exam 3	136	<.001	**0.64**
	Exam 4	139	0.011	0.25
Spring 2018 Stars & Galaxies	Exam 1	155	0.255	0.07
	Exam 2	153	0.965	0.003
	Exam 3	154	0.001	0.22

In all three courses, we see that for questions from exams that occurred within nine weeks of the final exam, there was a higher average normalized gain on questions previously considered as a group compared to questions previously considered individually. Adopting the convention that an effect size (Cohen's d; Cohen, 1969) in excess of 0.2 is significant, we note that the normalized gains for 8 of the 11 exams represent significant differences between the two sets of questions (individual and group) when revisited in the final exam. For questions derived from exams 1 and 3 in the Spring 2018 Solar System Astronomy class, we note a medium effect size.

In the Spring 2018 Stars and Galaxies course (Figure 3c), students performed significantly better on both individual and group exam questions from the first exam, which took place thirteen weeks before the final exam. Also, for this course, there were similar average normalized gains for both question sets for the first two exams. Only the third exam showed a substantial difference between the group exam questions and the non-group exam questions, a difference that was accompanied by a significant Cohen's d effect size.

Looking at the other courses, we find that for the Fall 2017 Solar System (Figure 3a) course, student performance on those questions considered as a group increased significantly more on Exams 2 & 3, but not Exam 1. Taken alone, this result suggests that the learning benefits of collaborative exams are time dependent, perhaps lasting on the order of 6 to 8 weeks.

The results for the Spring 2018 courses, though, differ somewhat. For the Spring 2018 Solar System course (Figure 3b), we find the average normalized gains on the group question sets to be statistically greater than the non-group question sets for all exams except the second one. In that case, the normalized gain on the individual question subset was greater, with a significant effect size.

Qualitative Results

Overall, 42 (30.9%) students who answered the survey were currently taking the Solar System Astronomy class, while 61 (44.9%) were currently taking the Stars and Galaxies class. 33 (23.5%) of the students had experienced group work and collaborative exams in both of these classes.

Before taking their first two-stage collaborative exam, students had been working with their permanently assigned groups on several in-class activities, thus giving them experience working with their group members by the time of the first exam. Asked whether they found the group exam discussion to be beneficial to their understanding of the material, 99 respondents (72.8%) responded positively (agree or strongly agree). Similarly, 119 (87.5%) agreed or strongly agreed that their group worked as a team during the group portion of the exams. On the other hand, 30 respondents (22.1%) agreed or strongly agreed that their group was dominated by one or two individuals.

The majority of students perceived that their points of view were given consideration during group interactions. 76.5% of students disagreed or strongly disagreed to the statement, “My group members often dismiss what I have to say.” In fact, only 12 of 136 respondents (8.8%) agreed or strongly agreed with the statement.

It does not appear that students were content simply sitting back while the other members of their group wrestled with the material. 111 out of 136 students disagreed or strongly disagreed with the statement, “I rely on the fact that I will get credit for group work even if I don’t put forth much effort.” Again, a tiny minority (9 out of 136, or 6.6%) of students agreed or strongly agreed.

It appears that truly collaborative activities were seen as overwhelmingly positive learning experiences. In fact, the very act of communicating their understanding improved students’ perception of their own learning. Over 85% of respondents agreed or strongly agreed with the statement, “Explaining things to my group members helps me better understand the material.” See Figures 4 and 5 for the complete breakdown of responses to each of the survey questions.

Figure 4. Responses to pairs of survey questions. Red indicates a “positive” attitude, while blue indicates a “negative” attitude, regardless of the statement. For the top charts, dark red indicates “strongly agree,” but for the bottom charts, the color coding is inverted, with dark red indicating “strongly disagree.”

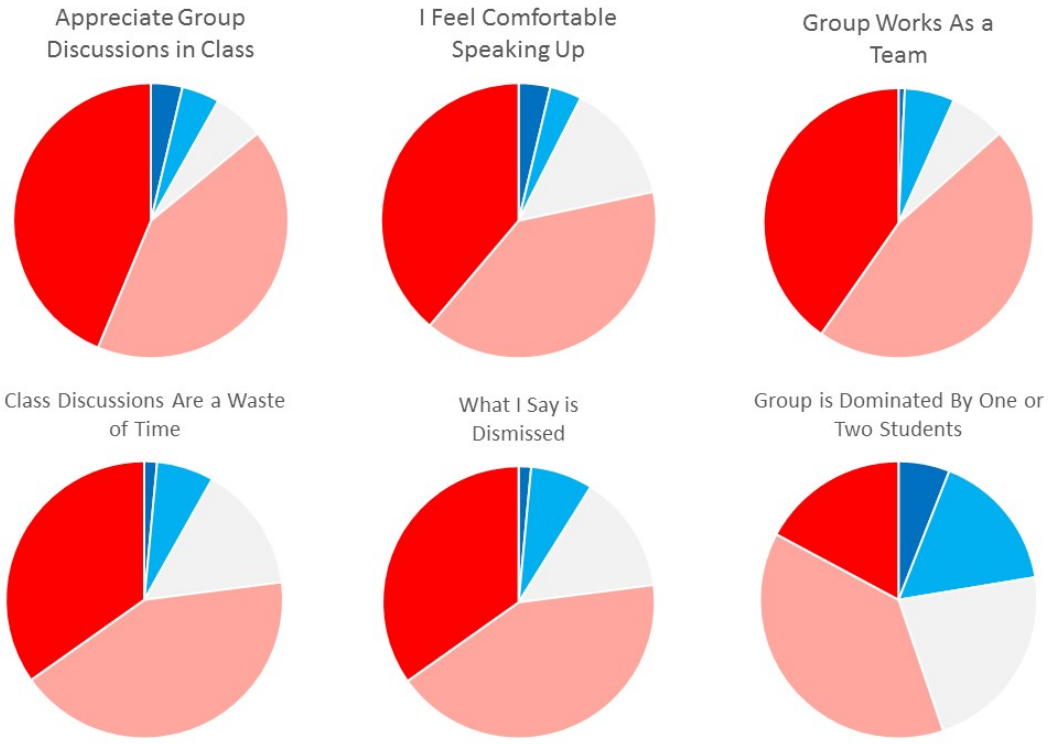
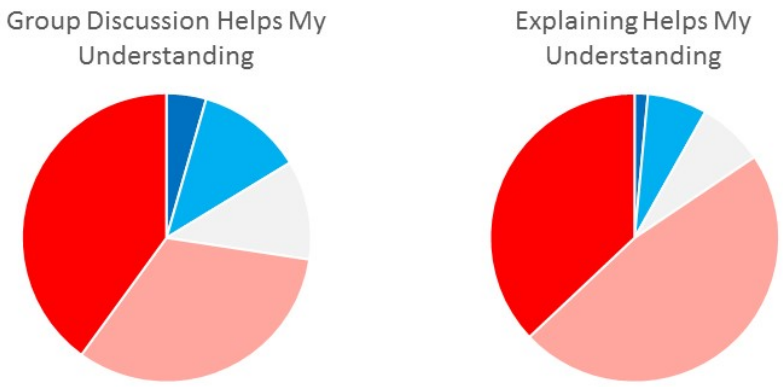


Figure 5. Responses to survey questions gauging student perception of the benefits of group interactions. Dark red indicates "strongly agree," followed by "somewhat agree" (pink), "neither agree nor disagree" (gray), "somewhat disagree" (light blue), and "strongly disagree" (dark blue).



In addition to the Likert survey data, we were afforded a wealth of comments (79 with any substance) from the respondents. In the most general sense, the comments can be categorized as positive group experiences, negative group experiences, and changes to individual behaviors as a result of the group experience.

Negative group experiences were cited in approximately one quarter of the comments. Among the most frequent negative comments were concerns about the negative impact of group work on individual grades, dominating members (one student even admitted to being "that group member"), condescending members, and frustration at the times that everyone in the entire group seemed equally confused.

Several students expressed the perception that the group portion of the exam, when combined with their individual exam grade, brought their exam averages down. "I don't like relying on others for my grade," stated one student. Another commented, "I hate the fact that I can fail a portion of the exam just because my group may not agree with me on what a correct answer choice is. I'd rather take my own test." Another student wrote, "When it comes to exams, I don't like [group work]. This class is a (sic) astronomy class, not a community appreciation class that makes me work better with others. My grade should be from my work and not effected (sic) by others." In a similar vein was this comment: "I have lost points on exams because my group is convinced that their answer is right when my answer was the right answer. I don't believe any other student should have a direct effect on my personal grade, especially when I know that not have (sic) the group would be better for me."

Interestingly, although the occasional student (approximately 3-5% per exam) earned a slightly higher score (0.5% to 3%) on the individual portion than on the combined exam grade for a particular exam, there were zero students at the end of the semester whose individual exam average was greater than the combined exam average. Thus, while some students perceived the group portion of the exam to be detrimental to their grade, it was in reality beneficial to everyone's grade.

Group dynamics – both positive and negative – were specifically mentioned in approximately 40% of the comments, with positive experiences slightly outnumbering the negative. Negative experiences included feeling ignored ("They do not listen even though I am right 99% of the time."), dominated ("The group 'leader' can persuade the entire group into believing that another answer is correct."), or taken advantage of. ("Some group members don't participate at all.").

Having experienced several in-class group activities and one group exam by the time of the survey, 22 students included comments suggesting that they had modified their study habits as a result of the group portion of the exam. Of those, 17 indicated that they had attended or were planning to attend study sessions with their group members – something they had not done before the first exam – to improve their aggregate understanding.

DISCUSSION

According to Laal & Ghodsi, 2012 (and references therein), in order for collaborative learning to be successful, there must be consistent use of relevant small-group skills, along with "frequent and regular group processing of current functioning to improve the group's future effectiveness." With this in mind, we organized our classes in such a way that group interactions were a daily occurrence in all the classes in an effort to maximize the groups' effectiveness. Many previous studies created groups only for the exams, leaving open the possibility that students had not become sufficiently comfortable in the group setting. We also ensured that the questions retested on the final exam were identical to the questions appearing on the original in-class exams. These two features set this study apart from previous investigations.

With this structure in place, we find that, with few exceptions, students in introductory astronomy courses for non-science majors perform better as a whole on multiple choice questions previously considered in a group setting than they do on questions previously answered individually. There are, however, some intriguing exceptions to this general finding.

Possible explanations for the exceptions include: 1) a limited time period over which students retain any increased understanding of a concept discussed collaboratively; 2) exam questions being so conceptually challenging that even groups struggled to identify the correct answer; and 3) repeated exposure to the same concepts on multiple exams, leading to enhanced learning regardless of encountering the question as a group or individually.

Addressing the first possibility, namely that the benefits of collaborative exams diminish with time, we find that our results for Fall 2017 Solar System and Spring 2018 Stars and Galaxies echo those from Ives (2014). However, it is possible that the pattern in normalized gains for the Spring 2018 Stars and Galaxies is a result of the format of the course itself. This course is designed such that the course content and the concepts involved continually build on the foundational material encountered in the first three weeks. The exams reflect this structure. The questions included on the first exam assessed understanding of this foundational material at the most basic level. That material was revisited with greater complexity and nuance throughout the semester, with frequent references to the basic concepts. Many questions on the second exam addressed these same concepts, but in greater depth. By the final exam, students had been assessed on these fundamental concepts on multiple occasions in multiple ways. This repeated exposure to the material might have led them to perform better on Exam 1 items regardless of their appearance on the collaborative portion of an exam. Because the third exam covered concepts of a different nature, results from this exam would have been unaffected by such continual reinforcement.

Indeed, a glance at the results for Spring 2018 Stars and Galaxies reveals the highest normalized gains for the material encountered on both individual and group questions derived from Exam 1, with slightly lower normalized gains for the questions derived from Exam 2. Consistent with Ives (2014), we see that individual questions derived from Exam 3 showed the lowest normalized gains, and the difference between the normalized gains for individual and group question sets from Exam 3 was the greatest, as was the effect size.

Exam 2 from the Spring 2018 Solar System course bears further scrutiny as well. In this case, the normalized gain for the group questions was appreciably lower than that for the individual-only questions, a feature not seen in Exams 1, 3, or 4. Delving deeper into the students' performance on the group questions, we note that their average performance on the group portion of the second exam was lower than their average performance on the group portions of the other exams, which were all consistent.

In particular, the average number of attempts (out of 5) to arrive at the correct answers on the IF-AT card for exams 1, 3, and 4 ranged from 1.29 – 1.38, whereas the average number of attempts to determine the correct answers for

exam 2 was 1.62, resulting in an average group score that was 7% lower than the other exams. For exams 1, 3, and 4, almost every group was able to answer successfully every question within their first or second attempt, but for five of the questions on Exam 2, roughly one-third of the groups required at least three attempts to ascertain the correct answer.

It is possible that the group questions on Exam 2 were more difficult than the group questions appearing on the other exams, and student groups were more often reaching the point where they were simply guessing at the answer. Rather than reinforce the concept being assessed, these group discussions might have inadvertently caused greater confusion. If this is the case, then this continued confusion would have negatively impacted their performance when revisiting those same questions on the final exam. This hypothesis will require further scrutiny.

The qualitative results indicate that, in general, the student groups were working well together and seeing the benefits of the group interactions. The qualitative results support this notion, as 87% of students agreed or strongly agreed with the statement "I appreciate group discussions in class," and 78% disagreed or strongly disagreed with the statement, "Class discussions are a waste of time." Not only did the vast majority of students agree or strongly agree that the group discussion helped their understanding (74%), they also agreed that the act of explaining material helped their understanding (86%).

Another reported benefit of collaborative learning is that it strengthens out-of-class bonds between group members. The qualitative comments seem to suggest that the unique experience of taking a portion of an exam – a major grade – as a group for the first time altered the approach to learning. After Exam 1, many students report that they were inspired to form study groups or to adjust their study habits in order to better prepare for the group discussions of the questions, as echoed in the following comments:

"As a group we will be getting together and studying to make sure we all understand the material."

"I am going to study with my group and question why they think the correct answer is ____ rather than ____."

"We will study together and ask each other questions."

Students who felt that their correct answers and explanations had been dismissed studied for later exams with an eye toward justifying their answers, as reflected in the following comments:

"I will study as if I'm preparing to explain my answers."

"I will study more to be able to back up why I believe something is right."

Another common thread in the comments that addressed study habits was the desire to perform better to avoid letting down their group members. "I plan on studying more so I'm not a burden to my group," wrote one student. Another commented, "We all want to try and carry our weight."

These comments and others suggest that several students became aware of the importance of both becoming a better team member and working to explain their understanding to the other group members. Whether these comments were written by those judged "lazy" or "freeloading" by other students is unknown. What is known is that the group portion of the exam encouraged the opposite behavior in at least some students.

CONCLUSIONS AND FUTURE WORK

For introductory astronomy courses for non-science majors at a regional public university, the collaborative portion of a two-stage exam appears to promote learning more than the individual portion. In addition, the collaborative exam experience appears to have improved student attitudes and study habits. When we compare student normalized gains on final exam questions that were included in a collaborative group exam to normalized gains on exam questions that

were individual-only, we find greater gains in general for questions encountered on the collaborative portions of the exams. These gains are frequently accompanied by a significant Cohen's d effect size.

The apparent increase in learning might be affected by a number of factors, such as diminished retention over time, overly-complex concepts being assessed, or increased time spent mastering concepts. Qualitatively, the opportunity to work as a group was reported by students to have a positive influence on their overall learning and their study habits.

What remains to be seen, however, are patterns in individual responses and group responses. Are group members always – or almost always – following one or two students who have established themselves as "the smart one?" Or are the discussions more egalitarian? Are students having rich, conceptual dialogs, or are they simply taking a vote? Do the group questions that challenge students the most see larger or smaller gains? Also of interest are those instances where a group collectively chooses the right answer on the first time, and yet no single group member chose the correct answer on the individual portion. How do groups talk themselves into the right answers in these cases?

To explore these questions and more will require a two-pronged approach. First, we need to take a more detailed look at the responses in the individual and collaborative portions of the exams. Second, we will obtain audio recordings of group discussions that could provide insights into the collective thought process. Those data, combined with the quantitative and qualitative data explored so far, should help us leverage the collaborative two-stage exam to maximize learning.

AUTHOR BIOGRAPHIES

Scott T. Miller is an Associate Professor of Physics at Sam Houston State University. He has taught introductory astronomy for non-science majors for over 16 years. His professional pursuits include astronomy education research, focusing on team-based learning skills as well as facilitating STEM education among pre-service and in-service science teachers. Please send all correspondence to Scott T. Miller: stm009@shsu.edu, Department of Physics, Sam Houston State University, Box 2267, Huntsville, Texas 77341

C. Renee James is a Professor of Physics at Sam Houston State University, where she has taught introductory astronomy for non-science majors since 1999. Her professional pursuits include tertiary astronomy education research, as well as science popularization and science communication. Email: phy_crj@shsu.edu

REFERENCES

Beaver, D. D. (2001). Reflections on scientific collaboration (and its study): past, present, and future. *Scientometrics*, *52*(3), 365-377.

Bruffee, K. A. (1987). The art of collaborative learning. *Change: The Magazine of Higher Learning*, *19*(2), 42-47.

Cohen, J. (1969) *Statistical Power Analysis for the Behavioral Sciences*. NY: Academic Press.

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*(3), 102-108.

Finelli, C.J., Inger, B., & Mesa, V. (2011) Student teams in the engineering classroom and beyond: setting up students for success. *CRLT Occasional Paper No. 29*, University of Michigan Center for Research on Learning and Teaching, Ann Arbor, Michigan. 12 pp.

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*, *43*(3), 83-91.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American journal of Physics*, *66*(1), 64-74.

Immediate Feedback Assessment Technique (http://www.epsteineducation.com)

Ives, J. (2014). Measuring the learning from two-stage collaborative group exams. *arXiv preprint arXiv:1407.6442*.

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

Knierim, K., Turner, H., & Davis, R. K. (2015). Two-stage exams improve student learning in an introductory geology course: Logistics, attendance, and grades. *Journal of Geoscience Education*, *63*(2), 157-164.

Laal, M., & Ghodsi, S. M. (2012). Benefits of collaborative learning. *Procedia-social and behavioral sciences*, *31*, 486-490.

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.
Meseke, C., Nafziger, R., and Meseke, J. (2010) Student attitudes, satisfaction, and learning in a collaborative testing environment. *Journal of Chiropractic Education*: Spring 2010, Vol. 24, No. 1, pp. 19-29.
Stearns, S.A. (1996). Collaborative exams as learning tools. *College Teaching*, Vol. 44, Iss. 3, p. 111.
Tonso, K. L. (2006). Teams that work: campus culture, engineer identity and social interactions. *Journal of Engineering Education*, 1(1), 1-13.
Yuretich, R. F., Khan, S. A., Leckie, R. M., & Clement, J. J. (2001). Active-learning methods to improve student performance and scientific interest in a large introductory oceanography course. *Journal of Geoscience Education*, *49*(2), 111-119.