THE OFFICIAL MAGAZINE OF THE OCEANOGRAPHY SOCIETY

Oceanography

Two-Stage Exams

A Powerful Tool for Reducing the Achievement Gap in Undergraduate Oceanography and Geology Classes

By Barbara C. Bruno, Jennifer Engels, Garrett Ito, Jeffrey Gillis-Davis, Henrietta Dulai, Glenn Carter, Charles Fletcher, and Daniela Böttjer-Wilson

ABSTRACT. As part of a school-wide course transformation project at the University of Hawai‘i to improve student learning and retention, multiple geology and oceanography instructors are introducing two-stage exams in their undergraduate courses. The first stage is the traditional, individual exam. The second stage is collaborative, in which groups of two to six students answer the same (or a subset of) questions posed during the first stage. We analyzed n = 289 scores on 14 two-stage exams in seven sections of five unique undergraduate courses taught by six different instructors. Two of the courses are categorized as oceanography and three as geology, although all courses cover both terrestrial and marine content. For each exam, the mean group score (stage two) exceeded the mean individual score (stage one), and all gains were statistically significant at $\alpha = 0.05$. Overall, the mean individual score was 73.2%, with a standard deviation of 15.0%. The mean group score was 89.6% with a standard deviation of 9.3%, reflecting an overall improvement and a narrowing of the achievement gap. Students who scored in the bottom quartile of the individual exam experienced the greatest improvement from individual to group (increase of 29.9 percentage points). This compares to a lower, but still statistically significant, increase of 5.5 percentage points for students in the top quartile. The majority (83%) of groups had a group score that exceeded the scores of all individuals in that group, which argues against the theory that the increased group score is due to group members simply copying answers from the top-performing individual in their group. A formal parametric (z) analysis reveals that the group scores are systematically higher than the maximum individual scores, indicative of a systematic (non-random) process. We interpret this process to be collaborative learning during the group stage of the exam. A cohort analysis reveals that groups containing all combinations of high- and low-performing students during stage one experience statistically significant mean gains in exam scores, and selecting groups to include a mix of high- and low-performing students is a highly effective way to proactively reduce the achievement gap.

INTRODUCTION

In the United States, fewer than 40% of students—and fewer than 20% of underrepresented minorities—who enter university with an interest in science, technology, engineering, and mathematics (STEM) finish with a STEM degree (PCAST, 2012). More than a decade of STEM education research shows that when active learning techniques are used in college courses, student achievement increases dramatically (Springer et al., 1999; Ruiz-Primo et al., 2011; Freeman et al., 2014). These gains are particularly pronounced for women (Lorenzo et al., 2006) and minorities (Haak et al., 2011; Snyder et al., 2016). In fact, statistically significant improvements are so great that if college classrooms were participating in controlled trials of medical interventions, lecturing control groups would be “stopped for benefit” (Pocock, 2006) and replaced by active learning techniques study-wide (Freeman et al., 2014). In other words, medical ethics would not permit students in the control group to continue to be subjected to the harmful effects of lecturing. Yet academic institutions have been slow to adopt active learning techniques, instead relying on the more traditional lecture format (Snyder et al., 2016).

In 2014, the School of Ocean and Earth Science and Technology (SOEST) at the University of Hawai‘i at Manoa formed an academic council to address issues such as undergraduate recruitment and retention. A key concern was the low four-year graduation rate. Of the 111 students who enrolled in a SOEST major from fall 2010 to spring 2012, only 32% (36) graduated with a SOEST degree within four years (Leona Anthony, Director of Student Services, SOEST, *pers. comm.*, 2016), comparable to the national average (PCAST, 2012).

In Hawai‘i, a majority of the state’s residents are ethnic “minorities” relative to the national ethnic landscape. These students are underrepresented at both undergraduate and graduate levels in SOEST (Bruno et al., 2016; University of Hawai‘i Institutional Research and Analysis Office, 2016). Our overarching goal is to attract a diversity of students to SOEST and create an environment in which they can thrive academically. As part of this effort, we embarked on a school-wide course transformation project. Thirty oceanography and geology instructors signed on to participate, and this strong show of support led to National Science Foundation funding. Here, we report on initial promising results from a specific intervention: the use of two-stage exams (Yuretich et al., 2001) in undergraduate oceanography and geology courses. The first stage is the traditional, individual exam. The second stage is collaborative, in which students work in small groups to answer the same (or a subset of) questions posed during the first stage.

This paper builds on previous work by (1) reporting results from multiple classes, instructors, and topics across an institution, (2) analyzing data by student achievement levels (quartile analysis), (3) addressing the question of whether the increased group scores reflect learning, and (4) examining the effect of mixing high- and low-performing students within a group.

BACKGROUND

Active learning is a general term that describes students doing anything other than passively listening to lectures (Bonwell and Eison, 1991). Examples of active learning instructional techniques include worksheets, think-pair-share, role-playing, group projects, concept maps, minute papers, discussions, and debates (Derek Bok Center, 2016; SERC, 2016). In a recent meta-analysis of 158 studies across a wide range of STEM disciplines, Freeman et al. (2014) showed that active learning instructional techniques increase mean exam performance by 6% and decrease failure rates by 35% (from 34% under traditional lecturing to 22% under active learning).

Importantly, active learning methods have been shown to help all students, but especially women, minorities, and lower-performing students. Haak et al. (2011) showed that active learning methods decrease the achievement gap by half between educationally advantaged students and disadvantaged students, with 77% of the latter being from underrepresented minority groups in their study. In addition, a study of Harvard undergraduate physics classes by Lorenzo et al. (2006) found that active learning techniques reduce achievement gaps between male and female students from 11% to being statistically indistinguishable.

Among the range of active learning techniques available to classrooms of different sizes and disciplines, peer collaboration has been shown to be particularly effective in improving undergraduate learning outcomes and persistence in majors (Lyle and Robinson, 2003; Tenney and Houck, 2003; Wamser, 2006; Arthurs and Templeton, 2009). Peer collaboration increases concept understanding (Lucas, 2009; Smith et al., 2009), retention (Deslauriers et al., 2011), participation (Lucas, 2009), comprehension of instructor explanations (Smith et al., 2011), and quantitative problem solving skills (Crouch and Mazur, 2001). Remarkably, peer collaboration alone has been shown to decrease course failure rates of minority students in biology classrooms from nearly 40% to 15%, effectively closing the achievement gap between minority and majority students (Snyder et al., 2016).

Collaborative (two-stage) exams have been successful in a variety of disciplines, including oceanography (Yuretich et al., 2001), geology (Knierim and Davis, 2015), biology (Leight et al., 2012), medicine (Lindsley et al., 2016), mechanical engineering (Fengler and Ostafichuk, 2015), and physics (Rieger and Heiner, 2014; Wieman et al., 2014). They have been shown to increase individual student knowledge and retention, for both high- and low-performing students (Gilley and Clarkston, 2014). Student survey data indicate they help students develop positive relationships with classmates (Sandahl, 2010), increase students' enjoyment of a course and reduce dropout rates (Stearns, 1996), and improve students' perception of an exam and motivation to study (Shindler, 2004). They may also help reduce students' test anxiety (Lusk and Conklin, 2003).

Two-stage exams combine active learning, peer collaboration, and assessment in an easy-to-implement format that can be used in classrooms of any size. Most of our instructors report that administering a two-stage exam requires little or no additional work, other than having to grade additional exam papers. Thus, two-stage exams can be an appealing first step ("low-hanging fruit") for faculty who want to introduce active learning into their classrooms, but do not have a lot of time to invest in learning new techniques or technologies.

DATA

Our data set comprises 289 student scores from 14 two-stage examinations given in seven sections of five unique undergraduate oceanography and geology courses taught by six different instructors in the University of Hawai'i system from 2012 to 2016 (Table 1). The five courses are: OCN 201 Science of the Sea; OCN 310 Global Environmental Change; GG 101 Dynamic Earth; GG 105 Voyage through the Solar System; and GG 106 Humans and the Environment.

Although two of the courses are classified as oceanography (OCN) and three as geology/geophysics (GG), all classes cover both terrestrial and marine content.

One course (OCN 201) was taught at Leeward Community College during the summer; the remaining six sections were held at the four-year research campus at Manoa during spring and fall semesters. One course (OCN 310) had a single, one-semester prerequisite; all others were introductory courses without any prerequisites. Data were provided to us secondarily by instructors (we did not interact with students) and anonymously (without student names or other identifying information). Thus, Institution Review Board (IRB) approval was not required.

All 14 examinations were given in two stages: an individual stage followed by a group stage. During the first stage, students took the exam individually (i.e., traditional method of exam taking). Then, the students turned in their exam papers and divided into groups. Groups were formed in various ways. In one class (OCN 201), student groups were assigned

TABLE 1. Summary of two-stage exam data set.

Exam ID[1]	Course Number[2]	Semester/ Year	Instructor ID[1]	Exam Type	n[3]	Group Setup[4]	Group Size	Groups Must Agree?[5]	Consult Notes? [6]	Grading Formula[7]
1	OCN 201	SUM 2016	1	Midterm	8	I-R	2	No	No	Max (I, G)
2	OCN 310	FA 2015	2	Midterm	21	S	5–6	Yes	No	Max (I, 0.85*I + 0.15*G)
3				Final	20	S	5	Yes		
4	GG 101	FA 2016	3	Midterm	73	S	3–4	Yes	Yes	Max (I, 0.5*I + 0.5*G)
5	GG 101	FA 2016	5	Midterm 1	26	S	2–3	Yes	No	Max (I, 0.8*I + 0.2*G)
6				Midterm 2	26	S	3–4	Yes		
7	GG 105	SPR 2012	6	Midterm 1	16	S	3–4	Yes		
8				Midterm 2	15	S	4	Yes	No	0.85*I + 0.15*G
9				Final	15	S	3–4	Yes		
10	GG 105	SPR 2013	6	Midterm 1	16	S	4	Yes	No	0.85*I + 0.15*G
11				Midterm 2	15	S	3–4	Yes		
12				Midterm 3	16	S	4	Yes		
13				Final	15	S	3–4	Yes		
14	GG 106	FA 2016	4	Midterm	7	I-M	2–3	Yes	No	Max (I, 0.8*I + 0.2*G)
				TOTAL	**289**					

[1] Exam ID and Instructor ID are unique identifiers, created for this study.

[2] Courses are Oceanography (OCN) and Geology and Geophysics (GG). All GG courses and OCN 201 are first-year courses with no prerequisites. OCN 310 is a second-year course with a one-semester prerequisite.

[3] Number of students who took both (individual and group) stages of exam.

[4] Groups are self-selected by students (S) or assigned by instructor (I). Instructor-selected groups can be random (I-R) or designed to mix achievement levels, based on prior performance (I-M).

[5] *Yes* means each group turned in a single exam paper, so students were forced to reach consensus. *No* means students turned in individual exam papers after group discussion.

[6] Were students allowed to consult notes during the exam?

[7] Instructor grading formulas used to calculate each student's total exam grade, based on their grades on Individual (I) and Group (G) stages. For students whose individual grade exceeded their group grade, some instructors just counted the higher individual grade as the total exam grade.

at random. In another class (GG 106), the instructor intentionally attempted to form groups of mixed achievement levels, based on performance on a prior assignment. In all other classes, students self-selected into groups. In all classes but one, the group members were required to turn in a single exam paper, forcing the group members to reach a consensus on each answer. In one class (OCN 201), students discussed the questions during the group stage but did not have to agree (each student turned in an exam paper before and after the group discussion).

In this study, group size varied from two to six students, with the instructors of very small classes (seven and eight students) having students work in groups of two. Class size ranged from seven to 73. Figure 1 and the online supplemental video show students working together during the group stage of a two-stage exam.

FIGURE 1. Students working together during the group stage of two-stage exams in various oceanography and geology classrooms at the University of Hawai'i.

METHODS

Standardizing the Data Set

This study is based on a retrospective analysis on the impact of two-stage exams; thus, not all courses were identical in approach. The instructors conducted the collaborative exams however they saw fit and collected data in slightly different ways. Therefore, an important first step was to standardize the data to the fullest extent possible before analysis.

This analysis consists of student data from two-stage exams, and only the 289 students who completed both the individual and the group stages were included. The original data set contained 293 students, but four (~1.4%) who only took the individual exam were removed. Those four students were absent during the regularly scheduled exam, and the makeup exam only included the individual portion.

The individual exams were not always identical to the group exam. For example, some individual exams had multiple choice and essay components, whereas the corresponding group exam only had multiple choice questions. In these cases, where scores on each component were available, we recalculated the individual exam scores based solely on the identical portions in order to make more direct comparisons. Also, some but not all exams offered extra credit (that is, bonus questions that allowed the maximum possible exam score to exceed 100%). In cases where scores on the extra credit portion were separately available, we excluded the extra credit and recalculated the individual exam scores based solely on the main portions of the exams. In all cases, scores were recalculated so the maximum possible score for each exam was 100%.

Analyzing the Data Set

The premise of this analysis is that the distribution of a collection of test scores reflects the "knowledge" of the students of that collection. As with all strategies of quantitative data analyses, there are shortcomings with this representation, but this is the basic premise with which we proceed.

Analysis 1. Group vs. Individual Scores (Analyzed for Each Exam)

For each exam separately, we computed basic descriptive statistics (mean and standard deviation) for the individual and the group data. The number of students who took each exam was small (n < 30) for all but one exam (Table 1). We therefore applied a paired two-tailed t-test to determine whether there was a statistically significant difference between the mean individual and the mean group scores, in other words, whether the "knowledge" of the students when taking the exam individually (stage one) differed from the "knowledge" of the same students when working collaboratively in groups (stage two). The significance level was set at α = 0.05.

Analysis 2. Group vs. Individual Scores (Entire Data Set, Analyzed by Quartile)

Here, we combined the 14 sets of exam scores into a single data set and plotted histograms of individual, group, and maximum individual scores (Figure 2). The maximum individual score is the highest stage one score in each group. To correct for variability among the different exams, we subtracted the mean individual grade for each exam from all of the grades (both individual and group) for that exam. This correction leads to a frequency distribution of individual scores that varies positively and negatively about a mean of zero, and causes all frequency distributions (individual, group, and maximum individual scores) to become more symmetric (Figure 3).

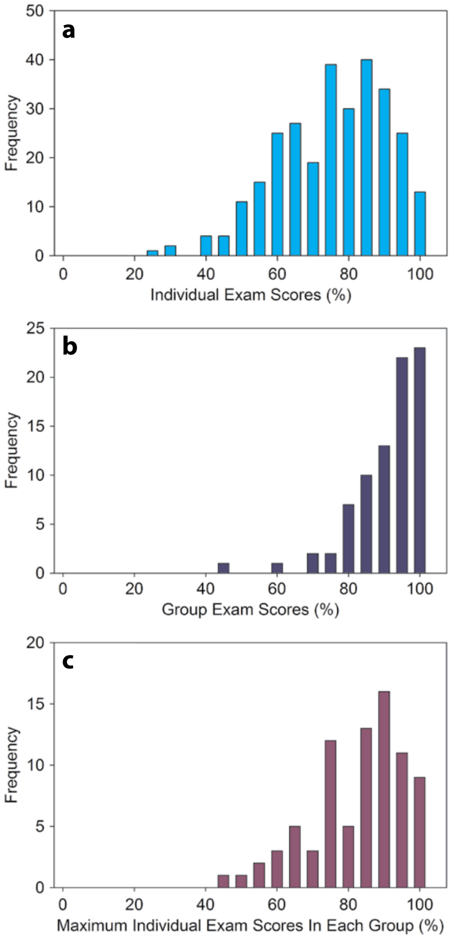

FIGURE 2. Frequency distribution of (a) individual (stage one) exam scores (%, n = 289), (b) group (stage two) exam scores (%; n = 81), and (c) maximum individual exam scores (%, n = 81). The maximum individual exam score is the highest stage one score within each group.

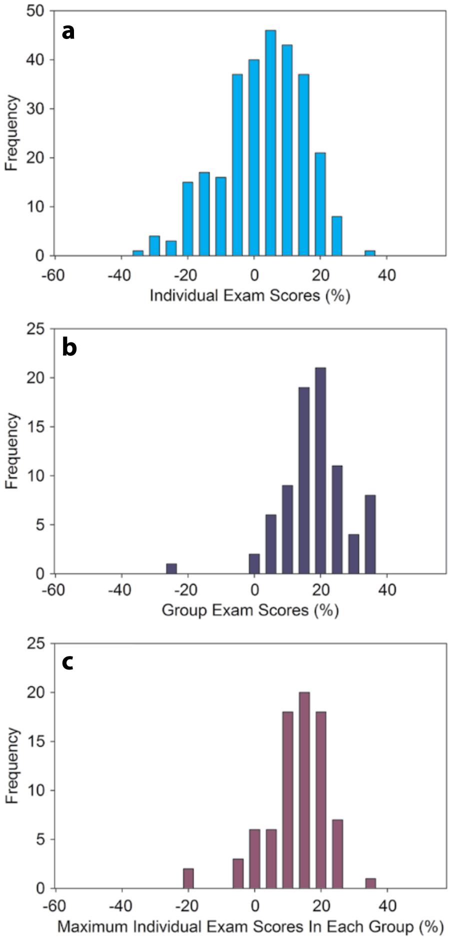

FIGURE 3. Frequency distribution of (a) individual (n = 289), (b) group (n = 81), and (c) maximum individual exam scores (%) in each group (n = 81), after correcting for variability among the different exams by subtracting the mean individual score for each exam from each individual and group score.

We then divided the de-meaned data into four quartiles based on the individual (stage one) scores. For each quartile, we calculated the mean individual score as well as the mean group (stage two) score of the students in those quartiles. This allows us to see how any improvements in scores from individual to group stages are distributed across student performance levels.

Unlike the individual exams, the quartiles have sufficiently large sample sizes to use a z-test for hypothesis testing. For each quartile, and for the data set as a whole, our null hypothesis was *that the students working collaboratively in groups during stage two performed with the same knowledge as they did when they performed individually during stage one* (i.e., their knowledge of the material on which they were being tested did not increase during the completion of the group exam). Thus, the appropriate standardization for the mean of the group scores represented in a given quartile is

$$z = \frac{\bar{x}_g - \bar{x}_i}{\sigma_i / \sqrt{n}}, \qquad (1)$$

where $\bar{x}_g$ is the mean of all group scores, $\bar{x}_i$ is the mean of all individual scores, σ_i is the standard deviation of the individual scores, and n is the number of groups. The numerator is the mean difference between individual and group scores. The denominator is an estimate of the standard error of $\bar{x}_g$, assuming the group scores come from the same knowledge that produced the individual scores. The z values were then converted to p-values of a normal distribution.

To determine whether any statistically significant differences are also meaningful in practice, we calculated the effect size for each quartile, and for the data set as a whole, using Cohen's d value (Cohen, 1988). For paired samples,

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}, \qquad (2)$$

where $\bar{x}_1$ and $\bar{x}_2$ are the sample means (in this case, means of individual and group grades), and s is the standard deviation of the differences. Cohen's d values are categorized as small (0.2), medium (0.5), large (0.8), very large (1.2), and huge (2.0) (Cohen, 1988; Sawilowsky, 2009).

Analysis 3. Group vs. Maximum Individual Scores (Entire Data Set, Analyzed by Group)

To evaluate the influence of having a high performing student in the group, scores of individuals within each group were analyzed. In particular, we were interested in testing the null hypothesis that *the students working collaboratively in groups during stage two performed with the same knowledge as did the highest performing student in each group during stage one* (i.e., the group score rose to match the individual grade of the highest performer within that group). Analogous to Analysis 2, we computed z of the mean of the group scores of the entire data set by assuming the group knowledge equals the knowledge of the highest performing student in each group,

$$z = \frac{\bar{x}_g - \bar{x}_{max}}{\sigma_{max} / \sqrt{n}}, \qquad (3)$$

where $\bar{x}_g$ is the mean of all group scores, $\bar{x}_{max}$ is the mean of the highest individual scores in each group, σ_{max} is the standard deviation of the highest individual scores, and n is the total number of groups. The denominator is an estimate of the standard error of $\bar{x}_g$, assuming the group scores come from the same knowledge that produced the maximum individual scores. Again, the z values were then converted to p-values of a normal distribution.

Analysis 4. Group vs. Individual Scores (Entire Data Set, Analyzed by Achievement Cohorts)

Finally, we were interested in seeing whether the mix of student achievement levels within a group correlated with the mean difference between individual and group scores. We divided the students into two broad categories: students who placed in quartiles 1 and 2 during stage one (L, or low individual scorers) vs. those who placed in quartiles 3 and 4 (H, high individual scorers). We divided the 81 groups into three categories: groups comprised entirely of low individual scorers (l, low groups), entirely of high individual scorers (h, high groups), and both low and high individual scorers (m, mixed groups). Thus, there were four possible achievement cohorts: Ll, Lm, Hh, and Hm. (By definition, there can be no Lh or Hl.) For example, Ll comprised students who performed in quartile 1 or 2 on the individual exam and whose fellow group members all placed in quartile 1 or 2 during stage one as well.

For each cohort, we calculated the mean individual score as well as the mean group score of the students in that cohort. Because some cohorts have small sample sizes (minimum n = 26), the statistical significance of any differences between mean individual and group scores was determined using a paired, two-tailed t-test. We then compared student achievement across cohorts. Comparing Ll vs. Lm (or equivalently Hh vs. Hm) allowed us to determine whether there was any correlation between the mix of student achievement levels within a group and any mean differences between individual and group scores. Comparing Ll and Hh allowed us to compare achievement levels of low vs. high individual scorers when placed in groups of students who performed similarly during stage one. Because the data among cohorts are uncorrelated, inter-cohort comparisons were done with an unpaired, two-tailed t-test.

RESULTS

Analysis 1. Group vs. Individual Scores (Analyzed for Each Exam)

A total of 289 pairs of scores from 14 midterm and final examinations were analyzed. Table 2 presents a summary of results. For each exam, the mean group score exceeded the mean individual score. Overall, across all exams, the mean individual score was 73.2% (standard deviation = 15.0%) and the mean group score was 89.6% (standard deviation = 9.3%), which represents a gain of 16.4 percentage points. Figure 2 shows the frequency distributions of the scores of individual students (n = 289), groups (n = 81), and the highest performing individual in each group (n = 81). Individual scores have an approximately normal distribution, while the group and maximum individual scores show distinct skewness toward higher scores (negative skewness). The paired, two-tailed, *t*-test for differences within each exam revealed that all group gains in each of the 14 exams were significant at α = 0.05 (all $p < 0.015$).

Analysis 2. Group vs. Individual Scores (Entire Data Set, Analyzed by Quartile)

Table 3 summarizes the results of the test for significance of the difference between the group and individual scores within each quartile, and for the entire data set. This analysis revealed that students in all quartiles—even the highest—showed a statistically significant mean improvement (all $p < 0.0001$) from the individual to the group stage. That is, all quartiles showed statistically significant differences between group and individual knowledge. However, this gain was not

TABLE 2. Individual vs. group scores (%) on two-stage exams.

Exam ID	Mean Individual Score[1]	Mean Group Score[1]	Mean Difference[2]	Standard Deviation Individual[3]	Standard Deviation Group[3]	*t*-statistic[4]	Critical *t*-value[5]	*p*-value[6]
1	67.5	77.2	9.7	8.6	7.4	−4.95	2.36	0.0017
2	69.4	84.0	14.6	15.5	6.7	−4.98	2.09	0.0001
3	83.4	94.8	11.4	8.6	2.9	−6.64	2.09	<0.0001
4	61.1	85.3	24.2	14.0	9.6	−17.75	1.99	<0.0001
5	70.1	85.5	15.4	15.5	14.2	−3.39	2.45	0.0148
6	80.4	90.6	10.2	10.8	7.1	−5.72	2.06	<0.0001
7	78.7	95.6	16.9	13.7	5.3	−5.52	2.06	<0.0001
8	78.8	88.8	10.0	11.8	9.6	−6.52	2.13	<0.0001
9	78.6	94.3	15.7	14.3	4.9	−4.17	2.14	0.0009
10	78.1	97.6	19.5	12.0	4.3	−6.90	2.13	<0.0001
11	84.2	97.9	13.7	13.0	1.6	−6.30	2.14	<0.0001
12	77.7	94.6	16.9	10.3	2.0	−4.72	2.13	0.0003
13	76.9	90.5	13.6	9.1	4.3	−3.90	2.14	0.0016
14	84.7	90.4	5.7	7.5	4.6	−4.92	2.14	0.0002
ALL	**73.2**	**89.6**	**16.4**	**15.0**	**9.3**	**−22.68**	**1.97**	**<0.0001**

[1] Mean Individual and Group Scores (%).
[2] Mean difference between Individual and Group scores. Positive number indicates higher group score.
[3] Standard deviation of the Individual and Group scores.
[4] *t*-statistic calculated using a two-tailed paired *t*-test.
[5] Critical *t*-value for statistical significance at α = 0.05.
[6] Probability value, all indicating statistical significance at α = 0.015 (all $p < 0.015$).

TABLE 3. Quartile analysis of de-meaned data.

Quartile	*n* (Individual)[1]	*n* (Group)[2]	Mean Individual Score[3]	Mean Group Score[4]	Mean Difference[5]	*z*[6]	*p*-value[7]	Standard Deviation Difference[8]	*d*[9]	Effect Size[10]
1	73	54	–16.7	13.3	29.9	30.5	<0.0001	12.0	2.5	Huge
2	73	54	–3.1	15.4	18.5	49.7	<0.0001	8.8	2.1	Huge
3	71	50	5.2	17.1	11.8	35.7	<0.0001	7.5	1.6	Very Large
4	72	49	14.9	20.4	5.5	8.7	<0.0001	8.2	0.7	Medium to Large
All Data	**289**	**81**	**0.0**	16.4	16.4	11.8	**<0.0001**	**12.3**	**1.3**	**Very Large**

[1] Number of Individuals in each quartile.
[2] Number of Groups in each quartile.
[3] Mean Individual exam score (%), after subtracting each exam's mean Individual score.
[4] Mean Group exam score (%), after subtracting each exam's mean Individual score.
[5] Mean difference between Individual and Group scores. Positive number indicates higher group score.
[6] *z*-statistic.
[7] Probability values, all indicating statistical significance at α = 0.0001 (all $p < 0.0001$).
[8] Standard deviation of the differences between the Individual and Group scores.
[9] Cohen's *d*, a measure of effect size, or practical significance (Cohen, 1988).
[10] Interpretation of Cohen's *d* (Sawilowsky, 2009).

evenly shared among student achievement levels. Instead, students who performed the lowest on the individual exam (quartile 1) experienced the greatest average grade improvement. Students in quartiles 1, 2, 3, and 4 improved an average of 29.9, 18.5, 11.8, and 5.5 percentage points, respectively.

To determine whether these statistically significant differences are practically meaningful, we calculated the effect size for each quartile, and for the data set as a whole, using Cohen's *d* value (Cohen, 1988). The lower two quartiles had huge effect sizes (*d* = 2.1–2.5). Quartiles 3 and 4 had very large (*d* = 1.6) and medium-to-large (*d* = 0.7) effect sizes, respectively. Overall, the effect size was very large (*d* = 1.3). In other words, students at all achievement levels benefited considerably from two-stage exams.

Analysis 3. Group vs. Maximum Individual Scores (Entire Data Set, Analyzed by Group)

Our finding that students in the lower two quartiles showed the greatest grade improvement suggests that two-stage exams can be used to reduce the achievement gap between high- and low-performing students, provided that learning is taking place during the group stage of the exam. But how do we know that students are truly learning during the group exam, and not simply copying off a high individual performer (i.e., the "smart kid") in the group? To answer this question, we analyzed the exam data within each individual group. If students were essentially just copying from a high-performing individual, then the group score would be expected to equal or approximate the highest individual score in that group.

For each of the 81 groups in this data set, we compared the group score with the maximum individual score within that group. In 67 groups (83%), the group score exceeded the maximum individual score (and hence all individual scores) in that group (Table 4 and Figure 4). The *z*-value for the overall mean group score relative to the mean maximum individual score within that same group was 5.2—that is, the group scores are 5.2 standard deviations greater than would be expected if the group scores come only from the knowledge of the highest performing student in each group. The corresponding *p*-value of <0.0001 is the probability of the group scores arising only from the knowledge of the high-scoring students. Therefore, we reject our null hypothesis and conclude that the knowledge of the students working collaboratively in groups during stage two is distinct from and greater than the knowledge of even the best-performing students when taking the exams individually.

Analysis 4. Group vs. Individual Scores (Entire Data Set, Analyzed by Achievement Cohorts)

Students in each cohort experienced mean gains of 15.8 (Ll), 25.4 (Lm), 8.6 (Hm), and 12.7 (Hh) percentage points, respectively, from individual to group scores (Table 5a). Not surprisingly, the greatest mean gain (25.4 percentage points) was experienced by the Lm cohort: low individual scorers placed in groups with at least one high individual scorer. However, all gains were statistically significant (all $p < 0.0001$) and sizable, with effect sizes ranging from large to huge. Even high-performing students put in a group with students of mixed achievement levels experienced mean gains of 8.6 percentage points.

Table 5b summarizes the results of

TABLE 4. Intragroup comparison of group and maximum individual scores.

Exam ID	# Groups	# Group Scores > Maximum Individual Score	% Group Scores > Maximum Individual Score
1	4	4	100%
2	4	0	0%
3	4	4	100%
4	20	20	100%
5	3	3	100%
6	9	8	89%
7	8	5	63%
8	4	4	100%
9	4	3	75%
10	4	4	100%
11	4	4	100%
12	5	2	40%
13	4	3	75%
14	4	3	75%
Total	**81**	**67**	**83%**

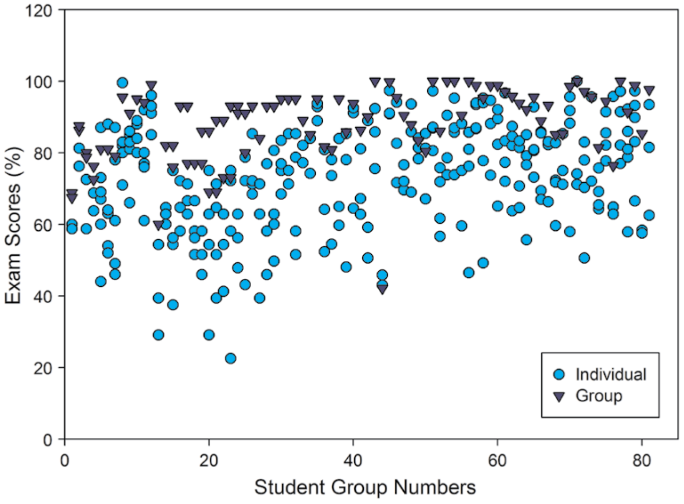

FIGURE 4. Scatterplot of individual (n = 289) and group (n = 81) exam scores (%) for the 14 analyzed exams. Scores for individual members of each group are placed in the same vertical plane as the group score.

an inter-cohort comparison (Lm vs. Ll; Hh vs. Hm; and Ll vs. Hh). Comparing Lm vs. Ll shows that low-performing students experienced a statistically significant (p <0.0001) benefit from being placed in a group with at least one high individual scorer, compared with a group of students of all low individual performers. High individual scorers received a statistically significant benefit (p <0.018) from being placed in a group of all high individual scorers, compared with a group of students of mixed achievement levels. In groups with more homogenous achievement levels (Hh and Ll), the low- and high-individual performers all experienced mean, statistically significant gains (Table 5a), and the difference in those gains is statistically indistinguishable (p = 0.22; Table 5b).

INTERPRETATION

The statistically significant difference between the group scores and the maximum individual scores (Analysis 3) indicates there is a systematic (non-random) process occurring during the group stage of the exam. But does this process necessarily entail learning? Is it possible, for example, that the group scores are higher simply because group members are copying correct answers from each other without learning? For this to happen, the students in each group would likely need to exhibit four key qualities: (1) systematic confidence for many correct answers; (2) systematic lack of confidence for many wrong answers; (3) systematic willingness to be vocal about their confidence or lack thereof; and (4) systematic willingness to copy other group members' answers, without understanding why those answers might be correct.

The first two qualities are characteristics of expert learners. Expert learners tend to be aware of the knowledge they do and do not possess (Ertmer and Newby, 1996). However, such experts tend to be unwilling to accept answers without understanding them. This leads us to a contradiction with the fourth quality, suggesting that the collaborative stage of the exam is a time during which students construct new knowledge through peer discussion. In other words, this suggests that collaborative learning is the systematic factor in raising group knowledge. This interpretation is consistent with Reiger and Heiner (2014), who found that only a small portion of groups relied on one person: 70% of groups engaged in discussion until consensus was reached.

This interpretation is reinforced by the compelling observation (Analysis 4) that groups containing only low-performing students during stage two experience gains that are statistically indistinguishable from the gains of groups containing only high-performing students (compare Hh and Ll, Table 5b). Stated another way, a student's prior "knowledge" going into the group exam does not enhance or diminish the efficacy of the group exam process: students of all achievement levels benefit.

This interpretation aligns with Gilley and Clarkston's (2014) assessment of learning retention. They administered a post-test as a surprise quiz three days after the group exam to assess knowledge retention, before giving the students any feedback on either their individual or group exams. Statistically significant improvements in individual student scores between the individual exams and the post-test indicate that (a) during the collaborative group exam, students acquired knowledge that they did not have previously; (b) learning was sufficient for students to retain this knowledge three days later; and (c) all students benefited,

TABLE 5. Intra-cohort (a) and inter-cohort (b) analyses of de-meaned data.

(a) INTRA-COHORT ANALYSIS									
Cohort[1]	*n*[2]	Mean Individual Score[3]	Mean Group Score[4]	Mean Gain[5]	Standard Deviation Difference[6]	*t*-statistic[7]	*p*-value[8]	*d*[9]	Effect Size[10]
Ll	32	–12.4	3.4	15.8	10.6	8.47	<0.0001	1.5	Very Large to Huge
Lm	114	–9.2	16.2	25.4	11.2	24.19	<0.0001	2.3	Huge
Hm	117	9.7	18.3	8.6	8.0	11.54	<0.0001	1.1	Large to Very large
Hh	26	11.9	24.6	12.7	7.8	8.31	<0.0001	1.6	Very Large to Huge

(b) INTER-COHORT ANALYSIS			
Cohorts' Mean Gains Compared[11]		*t*-statistic[12]	*p*-value[13]
Lm	Ll	4.32	<0.0001
25.4	15.8		
Hh	Hm	2.39	0.018
12.7	8.6		
Ll	Hh	1.23	0.22
15.8	12.7		

[1] Cohorts, determined by achievement level on individual exams. See text for details.
[2] Number of students per cohort.
[3] Mean Individual score (%), after subtracting each exam's mean Individual score.
[4] Mean Group score (%), after subtracting each exam's mean Individual score.
[5] Mean difference between Individual and Group scores. Positive number indicates higher group score.
[6] Standard deviation of the differences between the Individual and Group scores.
[7] *t*-statistic calculated using a two-tailed paired *t*-test.
[8] Probability value, all indicating statistical significance at $\alpha = 0.0001$ (all p <0.0001).
[9] Cohen's *d*, a measure of effect size, or practical significance (Cohen, 1988).
[10] Interpretation of Cohen's *d* (Sawilowsky, 2009).
[11] Cohorts' mean gains compared (Lm vs. Ll, Hh vs. Hm, Ll vs. Hh)
[12] *t*-statistic calculated using a two-tailed, unpaired *t*-test.
[13] Probability of the two cohorts' gains being equal.

with no significant difference between higher, middle, and lower performing students (Gilley and Clarkston, 2014).

Finally, instructor observations also suggest stage two of the exam may be an effective learning tool. Below are some (paraphrased) comments:

The two-stage exam works well. It gets students teaching each other—they use logic and critical thinking to formulate and defend arguments. I'd like to use this method of student peer-teaching more often, perhaps every class, because when I try to explain something, sometimes I'm "up here" and I don't always realize what they don't understand. Students can be good bridges to share knowledge. [Instructor 6]

I loved watching the group dynamics during this two-stage exam—especially when something clicked, and the group was able to agree on the answer. All of a sudden, the group members all broke out into a smile. How often does that happen during a test? [Faculty observer of Instructor 3] (see online supplemental video)

> "For faculty who are new to active learning classroom techniques, two-stage exams can be a time-efficient, high-reward entry point into peer collaboration."

I asked the class today what they thought of the two-stage exams. One student felt that having to explain an answer to her peers built her confidence in knowing she got it right during the individual part. Another student appreciated the opportunity to work with different students; through those experiences, she developed study partners as well as knowledge of who to work with during the final. When I asked the class whether they felt the two-stage exam format helped in learning, about a third of the class nodded vigorously. This whole experience makes me realize that group activities help students get to know each other and open up to each other, and it builds unity/camaraderie among the class. [Instructor 5]

The two-stage exam was a real success! The students were thrilled—they actively discussed the exam questions within their group and made links and connections to class activities and field trips. One of my Native Hawaiian students shared: "At the Halau [Hawaiian place of learning], we work and study as a group, not as individuals. We are a big family, looking out for each other and caring about each other. As a group, we rise together or fall together." So, the two-stage exam made a lot of sense to my students and me. [Instructor 1]

One student told me that she didn't understand a question on the individual exam but during the group exam she was paired with a high-achieving student who explained the question and the answer. I had a discussion with her after the exam and she was able to explain the answer correctly—clearly she now understands that material. [Instructor 4]

One student shared with me: *During the first [individual] part, I was really nervous so I did really badly. But then when we were working in groups, I felt more relaxed and was able to think more clearly. For one question, I was able to work through a problem and explain the answer to the group that none of us got right during the individual exam. It felt really good to figure out the answer and help my group understand it.* [Faculty observer of Instructor 2]

These qualitative observations support our interpretation that learning through peer discussion is occurring during the group stage of the exam. We received no comments indicating otherwise.

RECOMMENDATIONS

Based on both the quantitative data and qualitative observations, we highly recommend that instructors consider implementing two-stage exams in their undergraduate oceanography and geology courses. For faculty who are new to active learning classroom techniques, two-stage exams can be a time-efficient, high-reward entry point into peer collaboration. And, by making use of an already-scheduled assessment window, there is no class time "lost" to experimenting with new content delivery methods. To determine the exam score, the instructor can weigh the individual and group portions however they like: 75% and 25% (Yuretich et al., 2001) or 85% and 15% (Gilley and Clarkston, 2014) are typical. For students who perform better on the individual than the group stage, allowing the student to count the individual grade as their total exam grade can reduce any anxiety and/or ill feelings about having their grade negatively affected by the collaborative process. Previous studies indicate that group sizes of three to four or three to five students are ideal (Oakley et al., 2004; Carl Wieman Science Education Initiative, 2014).

Most attrition in STEM fields (PCAST, 2012) and SOEST (Leona Anthony, Director of Student Services, SOEST, *pers. comm.*, 2016) occurs during the completion of lower division coursework. Two-stage exams are a low-time-investment, high-impact way to help students develop positive relationships with classmates (Sandahl, 2010). They have also been shown to improve students' perceptions of an exam and their motivation to study (Shindler, 2004), increase students' enjoyment of a course, and, ultimately, reduce dropout rates (Stearns, 1996).

Class periods of at least an hour are ideal, but these exams can be given in shorter class periods with success. Most instructors found that they had to reduce the length of the exams by ~25% to fit into the allotted time, but determined the exams to be such helpful learning experiences for the students that they decided to increase the total number of exams during the semester. In this study, instructors reported modest increases in the time required to grade exam papers, but there was a clear consensus that the small additional time investment was justified by the robust learning gains.

The results of Analysis 4 suggest that the group stage of two-stage exams can be used as an efficient and powerful tool to proactively reduce achievement gaps between low- and high-performing students. Low-performing students achieve the highest gains of all cohorts when grouped with at least one high performing student, and the high-performing students still achieve statistically significant gains with large to very large effect sizes when grouped with low-performing students (Table 5a). We therefore advocate mixing different levels of students during the group exams to take full advantage of the opportunity for peer collaboration to improve the learning outcomes of all, but particularly those of lower-performing students.

CONCLUSIONS

We have found that two-stage exams can be an efficient, effective tool for reducing the achievement gap between high- and low-performing students in undergraduate oceanography and geology classes. In this study, each of the 14 exams showed a statistically significant gain between the individual and group stages of the exam (all p <0.015). Mean individual and group scores were 73.2% and 89.6%, respectively. A quartile analysis revealed that students at all achievement levels experienced statistically significant gains (all p <0.0001), with the lower-performing students showing the greatest gains. Students in the bottom quartile had a mean gain of 29.9 percentage points (Cohen's d = 2.5; huge effect size), compared with 5.5 percentage points (Cohen's d = 0.7; medium-to-large effect size) for top-quartile students. Overall, the effect size was very large (Cohen's d = 1.3). Analyzing the exam scores by group, we found that 83% of groups had a score that exceeded the scores of all individuals in that group. Our z-test results indicate that the group scores are extremely unlikely to arise from the highest individually performing students of the groups (p <0.0001). We interpret this finding to mean that knowledge is gained during the collaborative stage of the exam (as opposed to group members just copying answers from the top-performing individual). This interpretation is further strengthened by the observation that groups containing only students of similar skill levels to each other achieve gains that are statistically indistinguishable between high and low performers. We conclude that two-stage exams benefit all students, but can be particularly beneficial for low-performing students if they are grouped with higher-performing students.

FUTURE WORK

This analysis provides strong evidence that two-stage exams promote learning. But what type(s) of two-stage exams maximize learning? For example: (1) Is there a correlation between instructor grading formula and mean gains? and (2) What happens if students do not have to agree on an answer, but simply have to discuss the questions and then have the option to revise their individual answers? Moreover, it would be interesting to examine questions such as (3) Do learning gains during the group stage correlate with student gender, culture, or race? and (4) How do students who performed poorly on the individual stage of the first exam perform on the individual stages of subsequent exams and in subsequent semesters? These questions will be the subject of future analyses, as will the incorporation of post-tests to measure knowledge retention, as recommended by Gilley and Clarkston (2014).

ONLINE SUPPLEMENTAL MATERIAL

A video of students working together during the group stage of a two-stage exam is available at https://youtu.be/a498AAS90Ew.

REFERENCES

Arthurs, L., and A. Templeton. 2009. Coupled collaborative in-class activities and individual follow-up homework promote interactive engagement and improve student learning outcomes in a college-level Environmental Geology course. *Journal of Geoscience Education* 57(5):356–371, https://doi.org/10.5408/1.3544287.

Bonwell, C.C., and J.A. Eison. 1991. *Active Learning: Creating Excitement in the Classroom*. ASHE-ERIC Higher Education Report No. 1, 121 pp.

Bruno, B.C., J.L.K. Wren, K. Noa, E.M. Wood-Charlson, J. Ayau, S.L. Soon, H. Needham, and C.A. Choy. 2016. Summer bridge program establishes nascent pipeline to expand and diversify Hawai'i's undergraduate geoscience enrollment. *Oceanography* 29(2):286–292, https://doi.org/10.5670/oceanog.2016.33.

Carl Wieman Science Education Initiative. 2014. Two-stage exams, http://www.cwsei.ubc.ca/resources/files/Two-stage_Exams.pdf.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, United Kingdom, 567 pp.

Crouch, C.H., and E. Mazur. 2001. Peer instruction: Ten years of experience and results. *American Journal of Physics* 69(9):970–977, https://doi.org/10.1119/1.1374249.

Derek Bok Center for Teaching and Learning at Harvard University. 2016. Active learning, http://bokcenter.harvard.edu/active-learning.

Deslauriers, L., E. Schelew, and C. Wieman. 2011. Improved learning in a large-enrollment physics class. *Science* 332(6031):862–864, https://doi.org/10.1126/science.1201783.

Ertmer, P.A., and T.J. Newby. 1996. The expert learner: Strategic, self-regulated, and reflective. *Instructional Science* 24:1–24, https://doi.org/10.1007/BF00156001.

Fengler, M., and P.M. Ostafichuk. 2015. Successes with two-stage exams in mechanical engineering. *Proceedings of the Canadian Engineering Education Association (CEEA15) Conference, McMaster University, May 31–June 3, 2015*.

Freeman, S., S.L. Eddy, M. McDonough, M.K. Smith, N. Okoroafor, H. Jordt, and M.P. Wenderoth. 2014. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences of the United States of America* 111(23):8,410–8,415, https://doi.org/10.1073/pnas.1319030111.

Gilley, B.H., and B. Clarkston. 2014. Collaborative testing: Evidence of learning in a controlled in-class study of undergraduate students. *Journal of College Science Teaching* 43(3):83–91.

Haak, D.C., J. HilleRisLambers, E. Pitre, and S. Freeman. 2011. Increased structure and active learning reduce the achievement gap in introductory biology. *Science* 332:1,213–1,216, https://doi.org/10.1126/science.1204820.

Knierim, H., and R.K. Davis. 2015. Two-stage exams improve student learning in an introductory geology course: Logistics, attendance, and grades. *Journal of Geoscience Education* 63:157–164, https://doi.org/10.5408/14-051.1.

Leight, H., C. Saunders, R. Calkins, and M. Withers. 2012. Collaborative testing improves performance but not content retention in a large-enrollment introductory biology class. *CBE – Life Sciences Education* 11:392–401, https://doi.org/10.1187/cbe.12-04-0048.

Lindsley, J.E., D.A. Morton, K. Pippitt, S. Lamb, and J.M. Colbert-Getz. 2016. The two-stage examination: A method to assess individual competence and collaborative problem solving in medical students. *Academic Medicine* 91(10):1,384–1,387, https://doi.org/10.1097/ACM.0000000000001185.

Lorenzo, M., C.H. Crouch, and E. Mazur. 2006. Reducing the gender gap in the physics classroom. *American Journal of Physics* 74(2):118–122, https://doi.org/10.1119/1.2162549.

Lucas, A. 2009. Using peer instruction and i-clickers to enhance student participation in calculus. *PRIMUS* 19(3):219–231, https://doi.org/10.1080/10511970701643970.

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

Lyle, K.S., and W.R. Robinson. 2003. A statistical evaluation: Peer-led team learning in an organic chemistry course. *Journal of Chemical Education* 80(2):132–134, https://doi.org/10.1021/ed080p132.

Oakley, B., R.M. Felder, R. Brent, and I. Elhajj. 2004. Turning student groups into effective teams. *Journal of Student Centered Learning* 2(1):9–34.

PCAST (President's Council of Advisors on Science and Technology). 2012. Engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics. 103 pp, https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/pcast-engage-to-excel-final_2-25-12.pdf.

Pocock, S.J. 2006. Current controversies in data monitoring for clinical trials. *Clinical Trials* 3(6):513–521, https://doi.org/10.1177/1740774506073467.

Rieger, G., and C. Heiner. 2014. Examinations that support collaborative learning: The students' perspective. *Journal of College Science Teaching* 43(4):41–47, https://doi.org/10.2505/4/jcst14_043_04_41.

Ruiz-Primo, M.A., D. Briggs, H. Iverson, R. Talbot, and L.A. Shepard. 2011. Impact of undergraduate science course innovations on learning. *Science* 331:1,269–1,270, https://doi.org/10.1126/science.1198976.

Sandahl, S.S. 2010. Collaborative testing as a learning strategy in nursing education. *Nursing Education Perspectives* 31(3):142–147.

Sawilowsky, S. 2009. New effect size rules of thumb. *Journal of Modern Applied Statistical Methods* 8(2):467–474.

SERC (Science Education Resource Center at Carleton College). 2016. Starting point: Teaching entry level geoscience—Active learning, http://serc.carleton.edu/introgeo/gallerywalk/active.html.

Shindler, J.V. 2004. "Greater than the sum of the parts?" Examining the soundness of collaborative exams in teacher education courses. *Innovative Higher Education* 28:273–283, https://doi.org/10.1023/B:IHIE.0000018910.08228.39.

Smith, M.K., W.B. Wood, W.K. Adams, C. Wieman, J.K. Knight, N. Guild, and T.T. Su. 2009. Why peer discussion improves student performance on in-class concept questions. *Science* 323:122–124, https://doi.org/10.1126/science.1165919.

Smith, M.K., W.B. Wood, K. Krauter, and J. Knight. 2011. Combining peer discussion with instructor explanation increases student learning from in-class concept questions. *CBE – Life Sciences Education* 10:55–63, https://doi.org/10.1187/cbe.10-08-0101.

Snyder, J.J., J.D. Sloane, R.D.P. Dunk, and J.R. Wiles. 2016. Peer-led team learning helps minority students succeed. *PLoS Biology* 14(3):e1002398, https://doi.org/10.1371/journal.pbio.1002398.

Springer, L., M.E. Stanne, and S.S. Donovan. 1999. Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology. *Review of Educational Research* 69(1):21–51.

Stearns, S.A. 1996. Collaborative exams and learning tools. *College Teaching* 44(3):111–112, https://doi.org/10.1080/87567555.1996.9925564.

Tenney, A., and B. Houck. 2003. Peer-led team learning in introductory biology and chemistry courses: A parallel approach. *Journal of Mathematical Sciences* 6:11–20.

University of Hawaiʻi Institutional Research and Analysis Office. 2016. Enrollment, https://www.hawaii.edu/institutionalresearch/enrReport.action?reportId=ENRT00.

Wamser, C.C. 2006. Peer-led team learning in organic chemistry: Effects on student performance, success, and persistence in the course. *Journal of Chemical Education* 83:1,562–1,566, https://doi.org/10.1021/ed083p1562.

Wieman, C., G.W. Rieger, and C.E. Heiner. 2014. Physics exams that promote collaborative learning. *The Physics Science Teacher* 52:51–53, https://doi.org/10.1119/1.4849159.

Yuretich, R.F., S.A. Khan, R.M. Leckie, and J.J. Clement. 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, https://doi.org/10.5408/1089-9995-49.2.111.

ACKNOWLEDGMENTS

This project is funded by Hawaiʻi Experimental Program to Stimulate Competitive Research (EPSCoR) and Improving Undergraduate STEM Education (IUSE) Geopaths. Support for the Hawaiʻi EPSCoR Program is provided by the National Science Foundation's Research Infrastructure Improvement (RII) Track-1: ʻIke Wai: Securing Hawaiʻi's Water Future Award #OIA-1557349. Support for IUSE Geopaths is provided by NSF/GEO #1565950. In addition, we gratefully acknowledge the efforts of Sarah Bean Sherman and Sara Harris (Carl Wieman Science Education Initiative, UBC), Kathie Kane (UH Center for Teaching Excellence), Rosie Alegado (UH Oceanography Department), Leona Anthony and Marcie Grabowski (SOEST Student Services), and numerous other faculty and staff who are transforming SOEST into a place of learning where diverse students can thrive. The final version of this manuscript benefited from input from three peer reviewers.

AUTHORS

Barbara C. Bruno (barb@hawaii.edu) is Specialist, Hawaiʻi Institute of Geophysics and Planetology and Graduate Faculty, Department of Oceanography; **Jennifer Engels** is Assistant Specialist, Hawaiʻi Institute of Geophysics and Planetology; **Garrett Ito** is Professor, Department of Geology & Geophysics; **Jeffrey Gillis-Davis** is Associate Researcher, Hawaiʻi Institute of Geophysics and Planetology; **Henrietta Dulai** is Associate Professor, Department of Geology & Geophysics; **Glenn Carter** is Associate Professor, Department of Oceanography; **Charles Fletcher** is Associate Dean for Academic Affairs and Professor, Department of Geology & Geophysics; **Daniela Böttjer-Wilson** is Microbial Oceanography Specialist, Department of Oceanography, and Educational Associate, Hawaiʻi Institute of Geophysics and Planetology; all at the School of Ocean and Earth Science and Technology (SOEST), University of Hawaiʻi, Honolulu, HI, USA.

ARTICLE CITATION

Bruno, B.C., J. Engels, G. Ito, J. Gillis-Davis, H. Dulai, G. Carter, C. Fletcher, and D. Böttjer-Wilson. 2017. Two-stage exams: A powerful tool for reducing the achievement gap in undergraduate oceanography and geology classes. *Oceanography* 30(2):198–208, https://doi.org/10.5670/oceanog.2017.241.