CBE—Life Sciences Education
Vol. 13, 469–477, Fall 2014

Article

Identifying Key Features of Effective Active Learning: The Effects of Writing and Peer Discussion

Debra L. Linton,* Wiline M. Pangle,* Kevin H. Wyatt,† Karli N. Powell,‡ and Rachel E. Sherwood§

*Department of Biology, Central Michigan University, Mount Pleasant, MI 48859; †Department of Biology, Ball State University, Muncie, IN 47306; ‡Mathematics Department, Linden High School, Linden, MI 48451; §Science Department, Garden City High School, Garden City, KS 67846

Submitted December 20, 2013; Revised May 5, 2014; Accepted May 23, 2014
Monitoring Editor: Jennifer Momsen

We investigated some of the key features of effective active learning by comparing the outcomes of three different methods of implementing active-learning exercises in a majors introductory biology course. Students completed activities in one of three treatments: discussion, writing, and discussion + writing. Treatments were rotated weekly between three sections taught by three different instructors in a full factorial design. The data set was analyzed by generalized linear mixed-effect models with three independent variables: student aptitude, treatment, and instructor, and three dependent (assessment) variables: change in score on pre- and postactivity clicker questions, and coding scores on in-class writing and exam essays. All independent variables had significant effects on student performance for at least one of the dependent variables. Students with higher aptitude scored higher on all assessments. Student scores were higher on exam essay questions when the activity was implemented with a writing component compared with peer discussion only. There was a significant effect of instructor, with instructors showing different degrees of effectiveness with active-learning techniques. We suggest that individual writing should be implemented as part of active learning whenever possible and that instructors may need training and practice to become effective with active learning.

INTRODUCTION

Research in science education has identified several effective student-centered pedagogical techniques that have become the cornerstone of national efforts to reform science teaching (see *Vision and Change* report, table 3.2 [American Association for the Advancement of Science, 2011]). Cooperative group–based active learning is one of the most commonly implemented of these techniques (Ruiz-Primo *et al.*, 2011). Cooperative group–based active learning has been tested repeatedly and has been shown to result in significant learning gains in many individual studies (e.g., Udovic *et al.*, 2002, Knight and Wood, 2005; Armstrong *et al.*, 2007; Freeman *et al.*, 2007). Similarly, meta-analyses of active-learning research (e.g., Hake, 1998; Springer *et al.*, 1999; Prince, 2004; Wood, 2009; Ruiz-Primo *et al.*, 2011) have consistently supported the conclusion that these techniques can be effective in increasing student learning. However, in a random sample of college biology courses, Andrews *et al.* (2011) found that active-learning instruction did not correlate with student learning gains. They pointed out that instructors of courses in which science education research was being conducted were often science education researchers with knowledge and pedagogical experiences that facilitated implementation of activities. This leads us to expect that instructors without this knowledge may not have the same success implementing these strategies. Identifying key features of effective active learning is an important step in the dissemination of reformed teaching to these instructors through peer-reviewed literature and professional development programs (e.g., Pfund *et al.*, 2009; Ebert-May *et al.*, 2011; D'Avanzo, 2013).

DOI: 10.1187/cbe.13-12-0242
Address correspondence to: Debra L. Linton (linto1dl@cmich.edu).

There is a wide range of different pedagogical techniques that come under the umbrella term *active learning*. Prince (2004) defined active learning as "the process of having students engage in some activity that forces them to reflect upon ideas and how they are using those ideas" (p. 160). The differences come when we begin to look at what "some activity" means and how best to have students interact with it. Some examples of the types of activities commonly being implemented are problem-based learning, case studies, simulations, role-playing, conceptually oriented tasks, cooperative learning, and inquiry-based projects (e.g., Prince, 2004; Michael, 2006; Ruiz-Primo *et al.*, 2011). We focused our research on the technique identified by Ruiz-Primo *et al.* (2011) as the most common technique present in the research literature, which they identified as "conceptually oriented tasks + collaborative learning." In this technique, students work in groups on a task that requires some application of concepts to a problem or question. Almost 50% of the studies included in the Ruiz-Primo meta-analysis reported on the use of this strategy. Their analysis showed an effect size of 0.46–0.54, with an effect size of 0.5 (half an SD) typically considered "moderate." Yet we know that not all instructors who try this technique are successful (Andrews *et al.*, 2011).

A practical definition of effective active learning can best be built through studies that target individual components of active-learning design and implementation, instead of the effect of active learning as a whole, to identify what makes an effective active-learning exercise. There are many possible variations in the way an activity can be implemented. For example, students can discuss activities in groups or complete them individually. Students may only discuss aspects of the activity with others or they may be asked to write about their understanding as part of the activity, either individually or with one person per team writing the group's answer. Clickers are sometimes used as part of the processing of an active-learning exercise. Clicker questions may be answered individually or discussed in a group, or group discussion can follow after individual answers. The instructor may explain the correct answer to the activity after the work is completed or the instructor may have individual groups share their answers with the class and ask other groups to critique their answers. There are many such variations, and each leads to a question that can be investigated to help us build a shared and evidence-based definition of which of these options is most effective.

Some investigators have begun to conduct this type of research on the effectiveness of different modes of implementation of specific active-learning techniques. For example, Smith *et al.* (2009) explored the effect of peer discussion in the context of cooperative group–based instruction. They showed that students learned from group discussion of clicker questions and were able to apply their learning to answer novel questions on the concepts discussed. In a follow-up study, Smith *et al.* (2011) compared three different modes of implementing peer discussion of clicker questions and found that a combination of peer discussion followed by instructor explanation provided greater learning gains than either alone. If the science education community confirms these results through continued study, then we could begin to build a shared definition that includes the idea that effective active learning should include peer discussion followed by instructor explanation.

Student writing is another common feature of active-learning exercises. While a writing component is often included with the purpose of providing formative assessment data to the instructor, the concept of "writing to learn" suggests that writing also helps increase students' comprehension of complex concepts (Rivard, 1994). Writing about a concept requires students to examine and organize their thinking and thereby facilitates making connections between concepts (Bangert-Drowns *et al.*, 2004). Writing also provides an opportunity for self-assessment and metacognition (Armstrong *et al.*, 2008), as a learner is confronted with his or her own ability or inability to clearly articulate the concepts needed to answer a complex question. Meta-analyses of writing-to-learn research conducted in science (Rivard, 1994; Reynolds *et al.*, 2012) and non-science (Bangert-Drowns *et al.*, 2004) classrooms conclude that writing can improve student learning when implemented effectively. However, some studies (e.g., Armstrong *et al.*, 2008; Fry and Villagomez, 2012) did not find any effect of writing on student learning. These contradictory results have led to recommendations that future research should focus on determining the most effective implementation strategies for writing within specific instructional contexts. Within the context of active learning, students are often required to write about their understanding of a concept, either individually or in teams. The time required to grade and perhaps provide comments on written responses from hundreds of students in a large class setting is daunting. In addition, the time spent on student writing during class reduces time available for other activities and content coverage. A better understanding of the effects of writing on learner-centered outcomes would provide useful information as to whether or not writing is an effective use of class time.

The goal of our research was to identify some key features of effective active learning. In this study, we focused on evaluating the effectiveness of two implementation options, specifically peer discussion and writing, using a full factorial experimental design.

METHODS

Experimental Design and Implementation

We implemented this research during one semester in three lecture sections of an introductory biology course for biology majors. At the start of the semester, there were ~140 students in each section. The three sections were each taught by a different instructor. Instructor 1 (D.L.) has 14 yr of previous experience teaching large introductory biology courses and is a science education researcher who has been implementing cooperative group–based active-learning exercises for 10 yr. Instructor 2 (W.P.) has 4 yr of previous experience of teaching experience in large introductory courses and has been implementing active learning for 4 yr. Instructor 3 (K.W.) was teaching a large introductory biology course for the first time and had no previous experience implementing active-learning techniques.

The instructors met weekly to plan instruction and standardize delivery as much as possible, and the same lecture materials, in-class activities, and assessments were used in all sections. Students were assigned to four-person cooperative groups that were maintained throughout the semester. This course met three times each week for 50 min. During

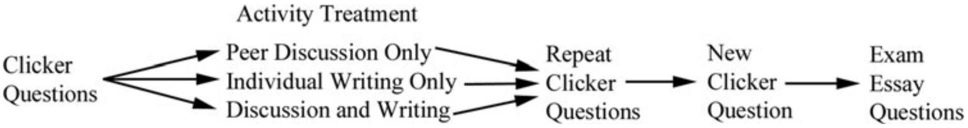

Figure 1. Experimental design.

the first two class meetings, the lesson consisted of lecture interspersed with multiple-choice questions that students answered using personal response systems (i.e., clickers), individually or with group discussion. One day each week, students completed an in-class activity that required application, analysis, or synthesis of the major concepts covered during two previous lectures. The experimental treatments were implemented within these activities (Figure 1). Each activity was preceded by one to three multiple-choice clicker questions dealing with key concepts for the activity. In all three treatments, students answered these questions without group discussion. The activity was then implemented. In one of the three sections, students completed the activity individually and wrote about their understanding of the concepts (writing-only treatment, WO). In the second section, students discussed the problem presented in the activity with their team, but did not write about their understanding (discussion-only treatment, DO). In the third section, students discussed the problem in their teams and then wrote individually about their understanding of the concepts (discussion and writing treatment, DW). The same clicker questions, followed by a new multiple-choice question, were then asked and answered by students without group discussion. After students turned in their activity worksheets, the entire activity was reviewed in a full-class discussion facilitated by the instructor. The three treatments were rotated by section each week, so each instructor implemented each of the three treatments a few times in his or her section. Students earned points based on the number of clicker questions answered correctly at the end of the activity and, in the WO and DW treatments, on the quality of their writing, to encourage students to make a good effort on all questions.

Most of the activities took ~30 min to implement (including the pre- and postclicker questions and full-class postprocessing). The DW treatment typically took a few minutes more than the DO and WO treatments. However, the three sections remained on pace with one another throughout the semester. The activities required students to perform one or more of the following tasks: making predictions, analyzing data, interpreting graphs, drawing models, explaining experimental results, and using evidence to support explanations of phenomena. We collected clicker performance data, activity writings, and exam writings for 10 different question sets based on these weekly activities, for the following concepts: the nature of matter, osmosis, transmission genetics, gene expression, natural selection, phylogenetics, community dynamics, ecosystems, carbon cycle, and global change. The essay questions on the exam were designed to be analogous (cover the same major concept and require the same skills) to the in-class activities. For example, for community dynamics, the in-class activity required students to make predictions, analyze data, and explain results based on Paine's (1966) classic *Pisaster* exclusion experiments from a rocky intertidal zone. On the midterm exam, students were asked to do the same with data from the Estes *et al*. (1998) sea otter study, and on the final exam, students predicted changes to a forest community based on proposed changes in population sizes of some species. A summary of each activity and the analogous exam questions are provided in the Supplemental Material. Many of these activities were based on published research studies (Spencer *et al.*, 1991; Ebert-May *et al.*, 2003; Stedman *et al.*, 2004; Winder and Schindler, 2004; Konopka *et al.*, 2009; Nowick *et al.*, 2009).

The three instructors met each week to debrief that week's activity and to plan for the following week's instruction. Based on the debriefing discussions, three of the activities were eliminated from the research data analysis; the phylogenetics and ecosystems activities were not implemented with enough standardization between sections to ensure there were no other factors influencing student learning, while the genetics activity did not effectively make use of writing and was more computational in nature. For the seven other activities, we analyzed clicker performance, in-class (activity) writings, and midterm and final exam data. The midterm and final exams consisted of a mix of multiple-choice and written assessments, with 40–50% of points on each exam coming from student writing that included essay questions analogous to those completed during the in-class activities. For natural selection, community dynamics, carbon cycle, and osmosis, we collected writing data from the activity, midterm exam, and final exam. For global change, nature of matter, and gene expression, we collected writing data from the activity and one midterm exam.

Data Analysis

Three hundred and forty-six students signed the consent forms to participate in the study and were included in the analyses. Students who were not present on activity days were removed from the analysis of exam questions, as they had not received the treatment for that concept.

Clicker question scores on the repeated questions were compared post- versus preactivity for each student and for each activity. If the student improved on the clicker questions from pre- to postactivity, his or her clicker performance was coded as "1" for improvement. If the student did not improve or performed worse on the postactivity questions, his or her performance was coded as "0." Sample sizes for clicker data, split by treatment and instructor, are shown in Table 1.

We coded all student writings for correct concepts based on coding rubrics developed for each assessment item by the research team. These rubrics were designed to parse out each individual correct concept that might be included in the students' writing. For example, the statement "Carbon dioxide from the atmosphere entered the plant through stomata in the leaves during photosynthesis" would be split into four concepts: 1) source of CO_2 is the atmosphere; 2) CO_2 enters the plant; 3) CO_2 enters through stomata in leaf; and 4) the process involved is photosynthesis. The "correct concepts" in the rubrics included not only statements of fact, but also

Table 1. Sample sizes for the clicker data and the exam writing data for all three instructors and across the three treatments[a]

	Instructor 1			Instructor 2			Instructor 3		
Treatment	DO	WO	DW	DO	WO	DW	DO	WO	DW
Number of observations	185	178	275	268	178	171	198	303	178
Number of students	107	102	99	99	100	98	116	116	109

[a] DO, discussion only; WO, writing only; DW, discussion and writing.

explanations of concepts, identifications of causal mechanisms, and statements of evidence used to support a conclusion. Therefore, this coding scheme allowed us to distinguish between different levels of completeness and complexity of students' answers on the essay questions.

A team of three raters coded the essays and disparities were settled by discussion. We calculated a total number of correct concepts for each student for each written assessment item. For concepts that were tested on both a midterm and final exam, the scores on the two essays were totaled to give a composite score on that concept. A total of 4477 essays were coded. Student writings were deidentified before analysis, so the evaluators could not determine to which treatment each writing belonged. Sample sizes for exam writings, split by treatment and instructor, are shown in Table 1. Sample sizes for in-class writings are shown in Table 2.

Because of differences in the complexity of the assessment questions (some of which were multipart), the total number of concepts in the coding rubric varied widely among questions, ranging from seven (the nature of matter) to 26 (community dynamics). To combine all essay data into a single data set for analysis, we normalized the essay scores based on the highest number of concepts included by any student on each essay question. For example, if the highest number of concepts included by any student on an essay question was 10, students who included 10 concepts would be assigned a score of 1, while students who included eight concepts would be assigned a score of 0.8.

We used generalized linear mixed-effect models (GLMM) with binomial distributions to analyze our data set. Generalized linear models are an extension of ordinary linear regressions. These models relate the responses of dependent variables to linear combinations of "predictor" independent variables. Whereas ordinary linear regressions assume a normal error distribution, the generalized linear models can take on a variety of other distributions. In our study, we used a binomial error distribution, which best fit the nature of our data. Our analyses are also termed "mixed" because they include both random and fixed factors; here, our random factor was the student unique identifier. We attempted to identify and account for as many independent variables that could potentially influence student performance in our study as possible. The GLMM analysis identifies which variables had a significant effect and also identifies any significant interactions between the independent variables.

In our analyses, student identification was entered as the random effect to avoid pseudoreplication (as one student could be represented up to seven times if he or she completed all seven activities). Three dependent variables were considered: 1) the improvement or lack of improvement on the clicker questions; 2) the standardized score for the essay writing during an in-class activity; and 3) the standardized score for the exam essay questions. For each dependent variable, we conducted a GLMM with three independent variables: 1) instructor (Instructor 1, Instructor 2, or Instructor 3); 2) treatment (DO, WO or DW); and 3) the average multiple-choice score of students on all exams (three midterm exams and one final exam) to account for aptitude and individual effort of students. The concepts assessed on the multiple-choice questions were not included as part of the activities, so student learning of these concepts should not have been influenced directly by the effects of the writing and discussion treatments. We prefer this measure of what we are calling "aptitude" over ACT scores or incoming grade point average, because it includes not only students' natural abilities but also allows us to account for variation due to study time, student motivation during the course, and other unmeasurable variables outside the classroom that might influence student performance specifically during the period of the course. Significance of the different independent variables was evaluated using the Wald χ^2 test.

Students' average performance on clicker questions across all activities was compared pre- versus postactivity (repeated questions only) using a Wilcoxon signed-rank test, as the data were not normally distributed. We also evaluated the effect of student aptitude (the average score obtained in all exams on the multiple-choice questions) on student clicker scores (the average postclicker score that included the new question added at the end of the activity) using a Spearman correlation.

Table 2. Sample sizes for the in-class writing data for all three instructors and across the two writing treatments[a]

	Instructor 1		Instructor 2		Instructor 3	
Treatment	WO	DW	WO	DW	WO	DW
Number of observations	175	237	171	162	273	172
Number of students	102	98	100	97	113	107

[a] WO, writing only; DW, discussion and writing.

All analyses were performed in the statistical software package R, version 2.1.0 (R Development Core Team, 2005), using two-tailed tests with $\alpha = 0.05$. The GLMMs were carried out using the R library MASS (glmmPQL function; Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993), while the Wald tests used the aod R library (wald.test function). Unless otherwise indicated, means ± SE are represented.

RESULTS

Clicker Question Performance

Students scored significantly higher on the postactivity clicker questions than on the preactivity questions ($V = 74{,}090$, $df = 1932$, $p < 0.0001$) in all treatments. However, this improvement was not significantly different between treatments (Figure 2; $\chi^2 = 2.8$, $df = 2$, $p = 0.24$). Students' average multiple-choice exam scores did not correlate significantly with improvement in clicker scores from pre- to postactivity ($\chi^2 = 1.7$, $df = 1$, $p = 0.19$). However, students' average multiple-choice exam grades were positively correlated with postactivity clicker scores ($r = 0.24$, $df = 1932$, $p < 0.0001$). Students under different instructors did not differ in their improvement on their clicker scores ($\chi^2 = 3.8$, $df = 2$, $p = 0.15$), nor were there treatment × instructor interactions (Figure 2; $\chi^2 = 4.1$, $df = 4$, $p = 0.39$).

In-Class Writing Data

Students' average multiple-choice exam scores were very strongly correlated with how well students performed in the writing during an activity, regardless of the treatment they received in lecture ($\chi^2 = 111.6$, $df = 1$, $p < 0.0001$). However, there were no differences between the WO and the DW treatments (Figure 3; $\chi^2 = 0.29$, $df = 2$, $p = 0.87$). There was a strong effect of instructor on students' performance during the in-class writing (Figure 3; $\chi^2 = 13.3$, $df = 2$, $p < 0.001$), with some instructors achieving higher student scores than other instructors regardless of the treatment and controlling for student aptitude. There were no significant treatment × instructor interactions (Figure 3; all p values >0.25).

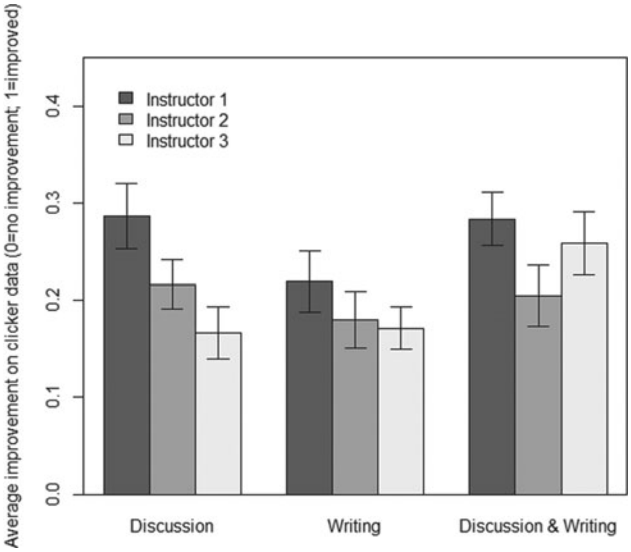

Figure 2. Effects of treatments and instructors on student improvement on clicker questions. Preactivity clicker scores were compared with postactivity scores and coded as "0" if students obtained the same or a lower clicker score after the activity or as "1" if students improved their clicker scores after the activity. Student scores are averaged by treatment received (WO, DO, or DW) and by instructors; error bars represent SEs. There were no significant effects of treatment ($p = 0.24$) or instructor ($p = 0.15$) on changes in clicker scores, and there was no treatment × instructor interaction ($p = 0.39$; see *Results* section for full GLMM results).

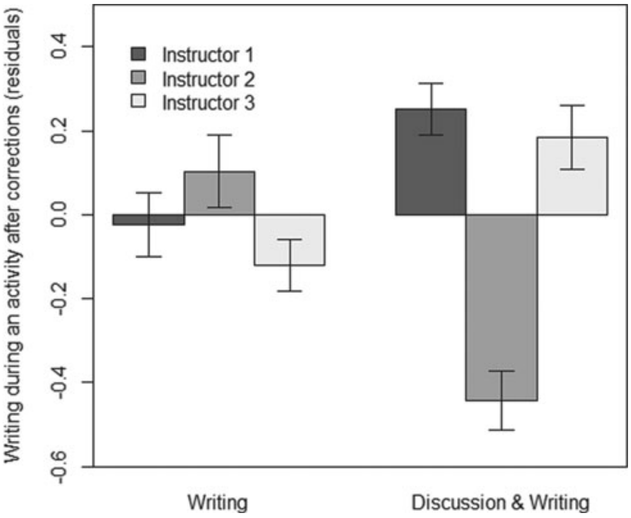

Figure 3. Effects of treatments and instructors on student writing scores during an in-class activity. Student writing samples were scored by the number of correct concepts and standardized across activities. Writing scores are corrected here for students' aptitude. Student scores are averaged by treatment received (WO or DW) and by instructor; error bars represent SEs. There is no significant effect of treatment ($p = 0.87$); however, there is an effect of instructor ($p < 0.001$). There are no significant treatment × instructor interactions ($p > 0.25$; see *Results* section for full GLMM results). In this figure, the negative average residuals indicate treatment × instructor combinations that resulted in lower averages on the in-class writing than predicted by the overall regression model that includes all data. Positive residuals indicate treatment × instructor interactions that resulted in higher averages than predicted by the model.

Exam Writing Data

Students' average multiple-choice exam scores were a very strong predictor of how well students performed in the writing during an exam, regardless of the treatment they received in lectures (Figure 4; $\chi^2 = 367.3$, $df = 1$, $p < 0.0001$). The treatments received during the lecture activity did have a significant effect on how well students performed on the exam writing (Figure 5; $\chi^2 = 7.2$, $df = 2$, $p = 0.027$). The WO and DW treatments resulted in higher performance on written exam components than DO. There was also a strong effect of instructor (Figure 5; $\chi^2 = 19.3$, $df = 2$, $p < 0.0001$) and a significant treatment × instructor interaction (Figure 5; $\chi^2 = 78.1$, $df = 4$, $p < 0.0001$), with some instructors achieving better student scores consistently within a certain treatment, which was not the same for each instructor. As we would expect, there was a very strong correlation between student

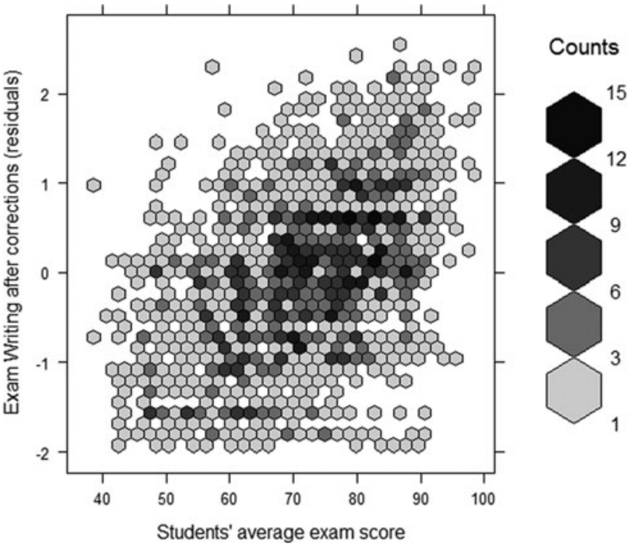

Figure 4. Effect of students' average exam score on the exam writing scores. Student writing samples during exams were scored by the number of correct concepts and standardized across activities. Writing scores are corrected here for treatment and instructor effects. For an accurate representation of the large number of data points, the plot area was divided in bins, which are then shaded based on the number of data points contained (dark bins contain up to 15 data points; light bins contain 1–3 data points). There is a significant positive correlation between multiple-choice exam scores and exam writing scores ($p < 0.0001$).

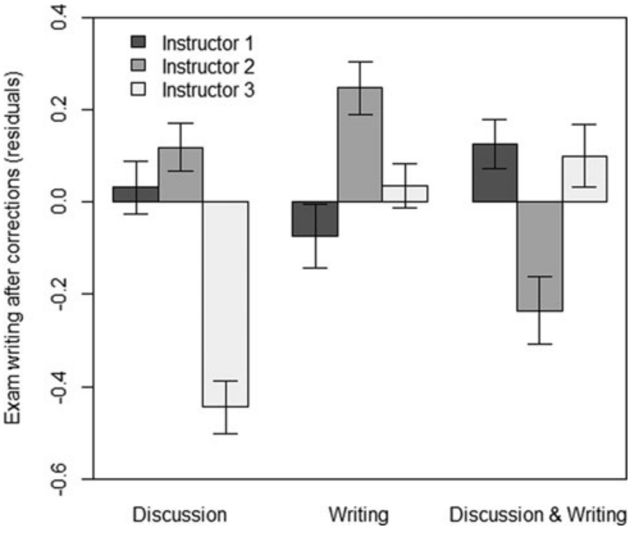

Figure 5. Effects of treatments and instructors on students' exam writing scores. Student exam writing samples were scored by the number of correct concepts and standardized across activities. Writing scores are corrected here for students' aptitude. Student scores are averaged by treatment received (DO, WO, or DW) and by instructors; error bars represent SEs. There are significant effects of treatment ($p = 0.027$) and instructor ($p < 0.0001$), as well as a significant treatment × instructor interaction ($p < 0.0001$; see *Results* section for full GLMM results). In this figure, the negative average residuals indicate treatment × instructor combinations that resulted in lower averages on the exam writing than predicted by the overall regression model that includes all data. Positive residuals indicate treatment × instructor interactions that resulted in higher averages than predicted by the model.

performance on the in-class writing and on the exam writing ($p < 0.0001$).

DISCUSSION

Student Effect

Students with higher aptitude performed better on all assessments, including the clicker questions. Students' average multiple-choice exam scores were highly correlated to their scores on the in-class and exam writing. This was expected, because the multiple-choice scores were included in the analysis as a measure of student aptitude to allow us to partially control for that variable in comparisons among experimental treatments. However, student multiple-choice scores did not correlate with student improvement on the clicker questions. Although the stronger students scored higher on the clicker questions, students at all levels improved equally (on average) as a result of experiencing the activities. Similarly, there were no treatment × student interactions found in any of the analyses. Neither stronger nor weaker students were advantaged or disadvantaged by any of the treatments when compared with the other students.

Treatment Effect

There were no treatment effects on clicker performance or in-class writing assignments but there were treatment effects on the exam writing. Both writing treatments (WO and DW) provided higher student performance than the DO treatment. The lack of treatment effect on the clicker scores suggests that improvement of clicker scores (from pre- to postactivity) was a result of participation in the activity, regardless of the method of implementation. The lack of significant difference between the WO and DW treatments on the in-class activity writing indicates that peer discussion did not improve student learning over that achieved by students thinking and writing individually. Instead, students were able to answer the activity questions as well on their own as they were after peer discussion.

Our results show that students who write about a concept perform better on subsequent writing-based assessments of that concept compared with students who only discuss the concept with peers in cooperative groups. We do not assume that this increased performance is a direct measure of increased student understanding of the concepts targeted by the activities. This increased performance could be due to increased understanding, but it could also be due to increased ability to communicate understanding in writing or increased retention of knowledge. The DW treatment did require more time than the DO or WO treatments; however, the fact that there was no significant difference found between the WO and DW treatments indicates that "time on task" was not a major factor.

The lower performance of the DO treatment suggests that writing is more important to student learning than peer discussion. With that said, we do not interpret our data as indicating that discussion is not important. In fact, other studies have reported gains in conceptual understanding following peer discussion (e.g., Smith *et al.*, 2009). We documented a similar trend during clicker activities, with students in the

DO treatment improving their scores on the clicker questions as a result of peer discussion of the activity. However, the DW and WO treatments led to similar improvements on the clicker questions and higher performance on the exam writings than did the DO treatment.

Research on "learning by explaining" has shown that explaining concepts can have a strong effect on learning and helps students make generalizations when presented with new applications of the same concepts (Williams *et al.*, 2010). Explaining has also been shown to facilitate conceptual change by encouraging metacognition (Williams *et al.*, 2010). There is a distinction made in this literature between explaining to self and explaining to others (Ploetzner *et al.*, 1999). A student writing individually is essentially explaining to self, while during peer discussion, students are explaining to others. Early research on learning by explaining hypothesized that explaining to others would lead to greater learning by the explainer than explaining to self, but subsequent studies have not found this to be the case (Ploetzner *et al.*, 1999). This suggests that peer discussion should be as effective as writing. All three treatments in our study required that students explain their understanding of the concepts. However, the mode of explanation was different in the three treatments. Students explained their understanding verbally, made written explanations, and made both verbal and written explanations. Based on the writing-to-learn literature (e.g., Rivard, 1994), we predicted that writing would require more careful organization of student thinking and, by doing so, might lead to greater understanding; our data support this prediction.

We have noted that, during a group discussion, it is rare that *each* student explains his or her understanding. Typically, one or two students attempt an explanation and the others agree, sometimes disagree and explain why, but often one or two team members do not speak during any given discussion (personal observation). Enforcing individual writing requires each student to explain his or her thinking. This could further explain why writing was more effective than peer discussion in this study.

Instructor Effect

The effect of instructor was strong for both the in-class and exam writings; however, we do not have sufficient data to infer what may have caused this difference. Previous research that included an instructor effect identified that both years of experience and training with active-learning techniques were important variables (Andrews *et al.*, 2011). There was a wide disparity among the three instructors in both overall number of years of teaching large introductory biology courses and experience with active learning. The less-experienced instructors in our study did have reduced success implementing some of the student-centered activities. However, we have no way of separating the effect of overall teaching experience and experience with active learning, as we did not have an experienced instructor in our study who was new to active learning. Therefore, our small sample size for this variable does not allow us to make inferences about this effect.

Treatment × Instructor Interaction

There was a significant treatment × instructor interaction for the exam writing data. Different instructors may have been more effective with different treatments. However, this apparent interaction may have been an artifact of the differences in the difficulty of different concepts or the difficulty of the exam questions on those concepts. In our data set, the exam writing scores for the gene expression activity were considerably lower than the scores for all other activities across all three treatments. The treatment × instructor interaction may have been influenced by this trend. Instructor 1 used the WO treatment for the gene expression activity, Instructor 2 used the DW treatment, and Instructor 3 used the DO treatment. These treatments were identified by the GLMM analyses to be the ones that were least successful for these instructors (Figure 5). While it is still possible that different instructors could be inherently more effective with some pedagogical techniques than with other techniques, our data do not provide rigorous evidence to support this idea.

CONCLUSIONS

Our results provide evidence that individual writing should be included as part of cooperative group–based active-learning exercises whenever possible. Although writing uses class time, this appears to be time well spent. Individual student writing not only provides formative assessment data but also promotes metacognition, as students are confronted with trying to organize the understanding of concepts, making connections, and justifying their thinking. A major assumption implicit in this type of research is the assumption that "effectiveness" is validly measured by student performance on course assessments. We consider the effectiveness of a technique to be determined by how well it facilitates student learning. A technique or activity is effective if it helps students understand the concepts being presented, and it is "more effective" if students understand better after experiencing this technique or activity than another with which it is being compared. We used performance on our assessments as a proxy measurement for students' understanding. Our results show that the writing treatments led to significantly higher student performance on our assessments than the discussion treatment.

Although our results indicate that peer discussion is not as effective as writing in facilitating student learning, we do not recommend that it be removed from active-learning exercises. In addition to the previous research cited earlier supporting the inclusion of peer discussion, there are other learning objectives that can be met using this technique. Collaboration and verbal communication skills are often objectives for introductory biology courses, and these objectives will not be met by student writing alone. Further research is needed to determine the most effective mix of discussion and writing.

The effect of instructor is a variable that should not be overlooked in national reform efforts. Although our data are limited in their ability to inform this question, the significant effect we found lends support to the idea that this may be a key variable to be addressed. There is a growing realization that effective dissemination of active-learning techniques is a bottleneck to the transformations called for in *Vision and Change* (Ebert-May *et al.* 2011; D'Avanzo, 2013). As a community, we can find ways to make these techniques accessible (note that we do not say "easily accessible") to any instructor willing to make the effort.

One limitation to our experimental design was that it did not allow us to analyze the effect (or effectiveness) of the different activities. While the activities we designed varied somewhat in their complexity, all activities required students to make an explanation of a biological phenomenon based on evidence, either experimental results or models that students developed to support their thinking. However, the much lower assessment scores on one of the concepts indicates that the activity associated with that concept did not meet the learning objectives, regardless of the implementation method. We suggest that more focus should be given to testing the effectiveness of specific activities that instructors design for different concepts and that the results be published for the community to share. In the same way that we would share a new technique in molecular biology, we should be able to publish procedures that our colleagues can follow (and practice and improve on) to achieve the desired results.

In conclusion, at the implementation level, we recommend the increased use of individual student writing during active-learning exercises. At the theoretical research level, we encourage more research into 1) the effect of "instructor" on the effectiveness of active learning and how to mitigate this effect and 2) the implementation strategies and types of activities that lead to the greatest student learning. At the practitioner research level, we call for increased rigorous testing and publication of specific active-learning exercises with detailed descriptions of the activities, evidence-based recommendations for implementation, and data on effectiveness. As we continue to study what makes an activity effective and to identify effective activities, we aim to make active learning more accessible to all instructors who are passionate about student learning.

REFERENCES

American Association for the Advancement of Science (2011). Vision and Change in Undergraduate Biology Education: A Call to Action, Washington, DC.

Andrews TM, Leonard MJ, Colgrove CA, Kalinowski ST (2011). Active learning not associated with student learning in a random sample of college biology courses. CBE Life Sci Educ *10*, 394–405.

Armstrong N, Chang S, Brickman M (2007). Cooperative learning in industrial-sized biology classes. CBE Life Sci Educ *11*, 17–25.

Armstrong NA, Wallace CS, Chang S (2008). Learning from writing in college biology. Res Sci Educ *38*, 483–499.

Bangert-Drowns RL, Hurley MM, Wilkinson B (2004). The effects of school-based writing-to-learn interventions on academic achievement: a meta-analysis. Rev Educ Res *74*, 29–58.

Breslow NE, Clayton DG (1993). Approximate inference in generalized linear mixed models. J Am Stat Assoc *88*, 9–25.

D'Avanzo C (2013). Post-vision and change: do we know how to change? CBE Life Sci Educ *12*, 373–382.

Ebert-May D, Batzli J, Lim H (2003). Disciplinary research strategies for assessment of learning. BioScience *53*, 1221–1228.

Ebert-May D, Derting TL, Hodder J, Momsen JL, Long TM, Jardeleza SE (2011). What we say is not what we do: effective evaluation of faculty development programs. BioScience *6*, 550–558.

Estes JA, Tinker MT, Williams TM, Doak DF (1998). Killer whale predation on sea otters linking oceanic and nearshore ecosystems. Science *282*, 473–476.

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

Fry SW, Villagomez A (2012). Writing to learn: benefits and limitations. Coll Teach *60*, 170–175.

Hake RR (1998). Interactive-engagement versus traditional methods: a six-thousand-student survey of mechanics test data for introductory physics courses. Am J Phys *66*, 64–74.

Knight JK, Wood WB (2005). Learning more by lecturing less. Cell Biol Educ *4*, 298–310.

Konopka G, Bomar JM, Winden K, Coppola G, Jonsson ZO, Gao F, Peng S, Preuss TM, Wohlschlegel JA, Geschwind DH (2009). Human-specific transcriptional regulation of CNS development gene FOXP2. Nature *462*, 213–217.

Michael J (2006). Where's the evidence that active learning works? Adv Physiol Educ *30*, 159–167.

Nowick K, Gernat T, Almaas E, Stubbs L (2009). Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain. Proc Natl Acad Sci USA *106*, 22358–22363.

Paine RT (1966). Food web complexity and species diversity. American Naturalist *100*, 65–75.

Pfund C *et al.* (2009). Summer institute to improve university science teaching. Science *324*, 470–471.

Ploetzner R, Dillenbourg P, Preier M, Traum D (1999). Learning by explaining to oneself and to others. In: Collaborative-Learning: Cognitive and Computational Approaches, ed. P Dillenbourg, Oxford, UK: Elsevier, 102–121.

Prince M (2004). Does active learning work? A review of the research. J Eng Educ *93*, 223–231.

R Development Core Team (2005). R: A Language and Environment for Statistical Computing, Vienna: R Foundation for Statistical Computing. www.R-project.org (accessed 2 December 2013).

Reynolds JA, Thaiss D, Katkin W, Thomson RJ (2012). Writing-to-learn in undergraduate science education: a community-based, conceptually driven approach. CBE Life Sci Educ *11*, 17–25.

Rivard LP (1994). A review of writing to learn in science: implications for practice and research. J Res Sci Teach *31*, 963–983.

Ruiz-Primo MA, Briggs D, Iverson H, Talbot R, Shepard LA (2011). Impact of undergraduate science course innovations on learning. Science *331*, 1269–1270.

Smith MK, Wood WB, Adams WK, Wieman C, Knight JK, Guild N, Su TT (2009). Why peer discussion improves student performance on in-class concept questions. Science *323*, 122–124.

Smith MK, Wood WB, Krauter K, Knight JK (2011). Combining peer discussion with instructor explanation increases student learning from in-class concept questions. CBE Life Sci Educ *10*, 55–63.

Spencer CN, McClelland BR, Stanford JA (1991). Shrimp stocking, salmon collapse, and eagle displacement. BioScience *44*, 14–12.

Springer L, Stanne M, Donovan S (1999). Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: a meta-analysis. Rev Educ Res *69*, 21–52.

Stedman HH, Kozyak, Nelson A, Thesier DM, Su LT, Low DW, Bridges CR, Shrager JB, Minugh-Purvis N, Mitchell MA (2004). Myosin gene mutation correlates with anatomical changes in the human lineage. Nature *428*, 415–418.

Udovic D, Morris D, Dickman A, Postelwait J, Weatherwax P (2002). Workshop biology: demonstrating the effectiveness of active learning in an introductory biology course. CBE Life Sci Educ *7*, 361–367.

Williams JJ, Lombrozo T, Rehder B (2010). Why does explaining help learning? Insight from an explanation impairment effect. In: Proceedings of the 32nd Annual Conference on the Cognitive Science Society, ed. S Ohlsson and R Catrambone, Austin, TX: Cognitive Science Society.

Winder M, Schindler DE (2004). Climate change uncouples trophic interactions in an aquatic ecosystem. Ecology *85*, 2100–2106.

Wolfinger R, O'Connell M (1993). Generalized linear mixed models: a pseudo-likelihood approach. J Statist Comput Simulation *48*, 233–243.

Wood WB (2009). Innovations in teaching undergraduate biology and why we need them. Annu Rev Cell Dev Biol *25*, 1–20.