CBE—Life Sciences Education
Vol. 10, 346–356, Winter 2011

Article

Assessment of Learning Gains Associated with Independent Exam Analysis in Introductory Biology

Adrienne E. Williams,* Nancy M. Aguilar-Roca,* Michelle Tsai, Matthew Wong, Marin Moravec Beaupré, and Diane K. O'Dowd

Department of Developmental and Cell Biology, University of California, Irvine, Irvine, CA 92697-1280

Submitted March 16, 2011; Revised September 17, 2011; Accepted October 11, 2011
Monitoring Editor: Diane Ebert-May

This study evaluates the impact of an independent postmidterm question analysis exercise on the ability of students to answer subsequent exam questions on the same topics. It was conducted in three sections (~400 students/section) of introductory biology. Graded midterms were returned electronically, and each student was assigned a subset of questions answered incorrectly by more than 40% of the class to analyze as homework. The majority of questions were at Bloom's application/analysis level; this exercise therefore emphasized learning at these higher levels of cognition. Students in each section answered final exam questions matched by topic to all homework questions, providing a within-class control group for each question. The percentage of students who correctly answered the matched final exam question was significantly higher ($p < 0.05$) in the Topic Analysis versus Control Analysis group for seven of 19 questions. We identified two factors that influenced activity effectiveness: 1) similarity in topic emphasis of the midterm–final exam question pair and 2) quality of the completed analysis homework. Our data suggest that this easy-to-implement exercise will be useful in large-enrollment classes to help students develop self-regulated learning skills. Additional strategies to help introductory students gain a broader understanding of topic areas are discussed.

INTRODUCTION

Given the rapidly expanding database of knowledge in the biological sciences, introductory-level courses in universities are now placing less emphasis on knowledge acquisition and more on developing students who are independent, analytical thinkers (Boud, 2000; Kitchen *et al.*, 2003; Hoskins and Stevens, 2009). Current theory and research indicates that this development is facilitated by formative assessment opportunities that allow students to evaluate their progress while learning (Sadler, 1989; Black and Wiliam, 1998; Wood, 2009). During the creation and piloting of an introductory biology class that included development of self-regulated learning as a goal, we adopted a number of formative strategies others have used effectively in large entry-level courses (Allen and Tanner, 2005; Knight and Wood, 2005; Caldwell, 2007). These included use of clickers and group activities in lecture, online quizzes prior to class, and interactive small-group discussion sections.

Previous studies have suggested that analysis of missed questions on summative midterm exams may also be a useful, formative assessment and learning opportunity for students (Boehm and Gland, 1991; Bolt-Lee and Foster, 2000; Ley and Young, 2001). In-class reviews of graded exams, led by faculty or conducted in student groups, have been shown to significantly improve performance on retakes of identical exams in small-enrollment classes (Wininger, 2005; Drouin, 2010). The ability to use an exam for self-regulated learning would be particularly helpful for students in large classes, where formative assessment opportunities are often limited (Nicol and Macfarlane-Dick, 2006). Therefore, in 2007, we provided students in our large introductory biology class (>400 students/section) with rapid and secure access to scanned PDFs of their graded midterm exams. Students were reminded that the final exam was cumulative, and they were encouraged to review their exams to "learn from their mistakes." Although 75% of the students rated their graded midterm exam as a

DOI: 10.1187/cbe.11-03-0025
*First two authors contributed equally to the study.
Address correspondence to: Diane O'Dowd (dkodowd@uci.edu).

valuable study aid, a preliminary analysis revealed no difference in final exam performance between students who did and did not download their midterm exam.

The inability of our students to use the graded exam as an effective learning tool is consistent with the majority being novice learners, who often require guidance in acquiring skills for self-regulated learning (Schunk, 1990; Ley and Young, 2001). Therefore, we developed an easy-to-administer "learn from exam" (LFE) homework activity to provide students with guidance for analyzing questions from the midterm exam. The exercise focused on questions answered incorrectly by a large percentage of the class. Most of these questions were ranked as application- or analysis-level questions on Bloom's rating scale (Bloom *et al.*, 1956; Crowe *et al.*, 2008), and the exercise therefore emphasized learning in the context of these higher levels of cognition.

Previous research using exams as formative assessment tools has focused on learning gains associated with retesting on identical questions (Wininger, 2005; Drouin, 2010). However, when writing a comprehensive final exam, most instructors do not use questions that are identical to those on the midterm. Thus, to evaluate the effectiveness of the exam analysis homework in helping students learn course material in an authentic situation, we assessed performance on a new set of final exam questions matched by topic to the homework questions. The level of similarity between the matched question pairs was typical of the range we use in developing questions for cumulative finals and for new exams each year to test our students' understanding on the same set of core topics. While two question pairs were identical, the other 17 had differences in scientific context, wording, format, and/or topic emphasis. Research suggests that novices, like our introductory biology students, may require many cues to successfully transfer information learned in one context and apply it to a new situation (Chi *et al.*, 1981; Barnett and Ceci, 2002). This predicts that the benefit of the exercise will be related to the degree of similarity between topic-matched question pairs.

The goal of this study was to determine whether the independent analysis of graded midterm questions by students would be an effective learning strategy. Our data demonstrate that this analysis activity can improve performance on topic-matched final exam questions requiring application skills, and even analysis skills, when the same aspect of the subject is emphasized. However, the data also suggest that the deeper level of learning required to successfully answer a topic-matched question with a new emphasis will require additional guidance. Approximately half of the students indicated they would be likely to apply a similar strategy in future classes, suggesting students will continue to use this approach for self-regulated learning.

MATERIALS AND METHODS

Class Format

This study was conducted in an introductory biology course for majors called Bio 93: DNA to Organisms at the University of California, Irvine (UCI), and included one section in Fall 2008 (A) and two sections in Fall 2009 (A and B). Class sizes ranged from 360 to 440 students. The classes were 10 wk long, with three 50-min lectures per week. Two back-to-back lecture sections (A and B) were team-taught by two faculty members. In addition to lecture, students were required to attend one 50-min discussion per week (30 students per discussion section; discussions were led by graduate students). Grades were assigned based on two in-class quizzes (12.5%), a single midterm (25%, given at the beginning of week 5), a cumulative final (47.5%, given in week 11), and formative activities, including online quizzes, clicker questions, and participation in the discussion section (15%). Both the midterm and final exams included a combination of multiple-choice and free-response questions.

Table 1. Demographic data for student participants[a]

	Fall 2008: Section A ($n = 432$)	Fall 2009: Sections A and B ($n = 795$)
Females	63	65
Biology majors	69	75
Freshmen	79	87
Ethnicity:		
East Asian	44	51
White/Caucasian	16	13
Chicano/Latino	12	10
Black/African American	2	3
Other/Decline to state	24	23

[a] All values are given as percentages.

Participants

All participants were undergraduate students enrolled in the courses, and the majority were freshman in their first quarter at the university. The demographic profile of the study participants is shown in Table 1. Students were given the opportunity to anonymously opt out of the study, and those under 18 were excluded. This protocol received Institutional Review Board approval, and student identifiers were replaced with randomly generated ID numbers before the student assignments were analyzed.

Protocol for Graded Exam Activity

We tested slightly different protocols in the 2 yr. The approaches are outlined in the following two sections, and flowcharts are shown in Supplemental Figure S1. Examples of completed homework from 2008 and 2009 are shown in Figure S2.

Discussion Training: Fall 2008. Training and homework submission occurred in small-group discussion sections. One of us (N.M.A.) delivered a mini-lecture (15 min) to each of the 15 discussion sections in the week following the midterm exam. The mini-lecture described how to analyze a question and employed an example midterm exam question answered incorrectly by the majority of the class. The instructor reviewed the lecture(s) from which the question was drawn, the general concept addressed, and the relevant information presented in the question stem. The instructor also led a discussion of the logic required to arrive at the correct answer, and why the other options were incorrect.

Four questions missed by more than 40% of the students were selected from the midterm. Six randomly selected discussion sections (140 students) were assigned two of these questions, and nine of the discussion sections (222 students) were assigned the other two. Students were instructed to write a short paragraph to answer the following questions:

1. What lectures did the information come from?
2. Why is the right answer correct?
3. Why are the wrong answers incorrect?
4. Why did I answer the question incorrectly, or for students who answered correctly, why might my classmates have missed the question?

Students returned their completed homework to their discussion teaching assistant (TA) the following week, along with a printout of one page from their graded exam to verify they had downloaded and looked at their exam. Completing the assignment was part of the student's discussion grade and worth < 0.5% of the total points for the class.

Lecture Training: Fall 2009. To increase efficiency in the second year, the presentation of the LFE training was moved to lecture, and the homework was submitted online. The lecture training covered the same points as the previous year but was done twice: at the end of week 2, after an in-class quiz, and again in week 5, after the midterm. Four questions missed by more than 40% of the students were selected from each midterm. Half of the students in each lecture section were randomly assigned two midterm questions to analyze for homework, and the other half of the students were assigned two different midterm questions. All students were also required to choose a third question to analyze from a list of four additional questions from each midterm. Each student downloaded the assignment as a PDF document from the course webpage, completed it by hand, and scanned or photographed their work for upload to an online drop box. Students were awarded 0.5% of the total points for the class for completing each assignment, but they were not given individual feedback.

Midterm–Final Exam Topic-Matched Question Pairs

The course instructor (D.K.O.D.) developed the questions for each midterm exam, and these were modified following pre-exam evaluation by course TAs, with the goal of having each question accurately test student mastery of specific learning goals for the course. The research associates (A.W. and N.M.A.) developed final exam questions that were topic-matched to selected midterm questions and that would require students to have understanding of the same topics to arrive at the correct answers. In 2008, there were four question pairs: four assigned homework questions from the midterm and four topic-matched final exam questions. In 2009, section A had four assigned pairs and four student-choice pairs, and section B had three assigned pairs and four student-choice pairs. The study-related questions on the cumulative final exams represented 7% of total points in 2008 (8/120) and 17% of total points in 2009A and B (16/95).

The average point-biserial coefficent of the study questions was 0.39 ± 0.01 (mean ± SEM; $n = 38$), indicating these questions enabled reliable discrimination between students with different abilities (Ding *et al.*, 2006). The Cronbach's alphas for the three final exams were 0.83 (Final 08), 0.85 (Final 09A), and 0.77 (Final 09B), indicating that the exams were reliable tools for evaluating student performance (Cronbach, 1951; De Champlain, 2008).

Each midterm and final exam study question was ranked on a Bloom's scale (Bloom *et al.*, 1956; Crowe *et al.*, 2008) by the course instructor (D.K.O.D.) and two research associates (A.W. and N.M.A.). The three worked independently using the following rubric:

- Knowledge: Requires memorization of material as presented in lecture. Answer choices generally do not have significant distractors. These are infrequent on our exams, and were not used as study questions.
- Comprehension: Requires understanding of concept or terms, often with integration of material from different lectures. Answer choices include significant distracters (six study questions).
- Application: Requires prediction of a most likely outcome given a new situation or perturbation of an already-discussed system (23 study questions).
- Analysis: Requires interpretation of a data set (graph, table, or figure) and selection of best conclusion (nine study questions).

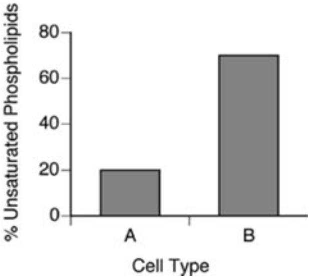
For all 38 questions, the Bloom's level was assigned based on agreement between at least two of three raters. Interrater reliability was high, with a Krippendorff's alpha of 0.88 (Freelon, 2010).

Each topic pair was also categorized on the basis of similarity between the midterm question and the final exam question, using the following rubric:

- Identical: Questions on both exams are exactly the same (two pairs).
- Same emphasis: The *same aspect of the topic* is emphasized, such as the effect of membrane fluidity on movement of a cell, but with differences in phrasing of stem and/or answer options, including changes in scientific context, organism, and/or cell type (nine pairs).
- Different emphasis: A *different aspect of the topic* is emphasized, such as the effect of membrane fluidity on movement of a cell versus the effect of membrane fluidity on movement of proteins within a cell membrane, and/or alteration in format (e.g., multiple-choice vs. short-answer; eight pairs).

All question pairs were categorized independently by six individuals: one instructor, two research associates, and three graduate student TAs associated with the class. While identical questions were in a clearly defined group, the distinction between Same Emphasis and Different Emphasis categories was more subjective, leading to some inter-rater variability (Krippendorff's alpha = 0.63). In 15 pairs, five of six raters agreed on one category. However, in four question pairs, two or three of the raters disagreed with the others about whether the pair was Different Emphasis or Same Emphasis. These questions of contested similarity were given their own Moderately Different Emphasis category. Examples of topic-matched question pairs are shown in Table 2.

Table 2. Samples of topic-matched midterm and final exam question pairs in the Same Emphasis, Moderately Different Emphasis, and Different Emphasis similarity categories[a,b]

<table>
<tr><th colspan="2">Assigned question pair 2: motor protein: Similarity category: Same Emphasis</th></tr>
<tr><td>Midterm 08 Bloom's level: application
An experimental drug unexpectedly reduced sperm motility in males. This drug most likely:
a. blocks dynein's ability to hydrolyze ATP*
b. inhibits extension of pseudopodia
c. blocks kinesin's ability to bind to microtubules
d. blocks assembly of actin filaments</td><td>Final 08 Bloom's level: application
Which of the following could prevent flagella movement in sperm?
a. A form of ATP that cannot be hydrolyzed*
b. A drug that blocks the actin binding site
c. A protein that stabilizes myosin
d. An enzyme that disrupts actin microfilaments
e. A drug that blocks the kinesin-binding site</td></tr>
<tr><th colspan="2">Student-choice question pair 15: cell signaling: Similarity category: Same Emphasis</th></tr>
<tr><td>Midterm 09A Bloom's level: application
Many chemotherapy drugs work by arresting the cell cycle. This prevents cancer cells from dividing, but also affects other healthy cells in the body. Two UCI researchers, Professors Longmuir and Robertson, have recently developed a method for delivering cancer-drug-carrying liposomes specifically to liver cells to treat liver cancer. This is possible because the liposomes:
a. contain a protein that recognizes liver cell specific DNA
b. synthesize and secrete specific tracking proteins
c. contain a protein that recognizes specific polysaccharides on liver cells*
d. contain a protein that recognizes specific enzymes on the smooth ER
e. contain a protein that recognizes specific phospholipids on liver cells</td><td>Final 09A Bloom's level: application
To create a drug-delivery liposome that would bind only to an ovarian cancer cell, the liposome should have:
a. a protein that recognizes ovarian cell DNA
b. the ability to synthesize and secrete specific tracking proteins
c. a membrane receptor that binds to polysaccharides on the surface of ovarian cancer cells*
d. a protein that recognizes ovary-specific enzymes on the smooth ER
e. a membrane receptor that binds to specific phospholipids on ovarian cancer cell membranes</td></tr>
<tr><th colspan="2">Student-choice question pair 17: membrane fluidity: Similarity category: Moderately Different Emphasis</th></tr>
<tr><td>Midterm 09A Bloom's level: analysis
Two types of cancer cells (A and B) found in the pancreas differ in the total concentration of phospholipids in their membranes that are unsaturated (see graph). Which of the following statements is most likely to be TRUE?
a. Cell type A shows evidence of less cholesterol in the membrane than B
b. Cell type B will move to other tissues faster than A*
c. Cell type A has more hydrocarbon tails with carbon = carbon double bonds than B
d. Cell type B has tyrosine kinase receptors that signal more slowly than A
e. Cell type A has more LDL in the membrane than B

</td><td>Final 09A Bloom's level: application
A researcher is testing the effects of garlic extract on cancer cells, and finds that cells exposed to allyl derivatives from garlic act as if they have an increased percentage of saturated phospholipids in the plasma membranes. Which of the following would you expect to see in these garlic-exposed cells compared to untreated cells?
a. Faster cell movement from one tissue to another
b. More membrane cholesterol
c. Less LDL in the membrane
d. Hydrocarbon tails with more carbon–carbon double bonds
e. Decreased lateral movement of proteins in plasma membrane*</td></tr>
<tr><th colspan="2">Assigned question pair 8: Na$^+$/K$^+$ pump; Similarity category: Different Emphasis</th></tr>
<tr><td>Midterm 09A Bloom's level: comprehension
If there is no ATP in the extracellular fluid how would this affect Na$^+$/K$^+$ pump function?
a. increase
b. decrease
c. no change*</td><td>Final 09A Bloom's level: comprehension
What are the two molecules from the intracellular fluid that bind to and are required for functioning of the Na$^+$/K$^+$ pump?
*(write-in answer) ATP and Na$^+$</td></tr>
</table>

[a] All comparisons were based on performance of Topic Analysis and Control Analysis groups on the final exam question in each pair.

[b] Asterisks indicate the correct answer.

Rubric for Assessing Quality of Analysis Homework

Students were not given any feedback on their homework submissions, but the quality of the responses was evaluated for the purposes of the study after final grades were submitted, using the following rubric:

- Strong: Analysis included clear explanation of the main biological concept and why the other answer options were incorrect.
- Weak: Analysis contained one or more misconceptions about the main biological concept and/or why other answer options were incorrect.

Assessment of Learning Gains

All comparisons were based on performance on final exam questions that were topic-matched to homework questions selected from the midterms. There were 19 question pairs total. Each student completed an analysis homework that contained only a subset of midterm questions, but answered final exam questions that were topic-matched to all midterm questions that had been assigned as homework. Therefore, within a class, each student was a member of a Topic Analysis group for final exam questions topic-matched to their homework. Each student was also a member of a Control Analysis group for final exam questions not matched to their homework assignment. Students who did not complete the homework were removed from further analysis.

For the assigned homework topics, the percentage of students who correctly answered the final exam question in the Topic Analysis group was compared with the percentage who answered correctly in the Control Analysis group. To evaluate the overall effect of the activity, we included students in the comparison regardless of whether or not they answered the original midterm question correctly.

For the student-choice topics, students were instructed to select a question they missed on the midterm from a list of four possible questions. This assessment therefore included only students who answered the midterm question incorrectly. The percent correct on the final exam in the Topic Analysis group was compared with the Control Analysis group.

The relationship between question-pair similarity and effectiveness of the exercise was determined by calculating the difference in percent correct on the final exam question: Topic Analysis (%) minus Control Analysis (%). Since this was evaluated for both assigned and student-choice topics, it included only students who answered the midterm question incorrectly.

Comparisons were also made within the Topic Analysis group between students turning in a strong versus weak exam analysis homework. This assessment included only students who answered the midterm question incorrectly.

GraphPad InStat software (version 3.1a, www.graphpad.com) was used for statistical analysis, and all comparisons were done using Fisher's exact tests.

Electronic Return of Graded Exams

Rapid Return is the online, electronic-document return system we helped develop and is currently in use campus-wide at UCI. After exam grading is complete, the exam booklets and scantron forms are picked up by UCI's Distribution and Document Management Services. All documents are scanned, and the PDFs are matched to student IDs. When the PDFs are available for posting (usually within 72 h of grading the exams), they are linked to the student's personal account in the course-management system. Faculty can also download the PDFs for permanent storage.

Survey

In 2009, after final grades were submitted to the registrar, all students received a request to complete an online anonymous survey that included questions about the exam analysis activities.

RESULTS

The majority of students completed the midterm exam analysis homework exercise in both years (83% in 2008; 92% in 2009).

Assigned Midterm Analysis Questions

In 2008, students were taught how to analyze an exam question in a mini-lecture given in small discussion sections the week after the midterm and then assigned analysis homework on two of four questions missed by more than 40% of the class on the midterm. One group was assigned an analysis homework activity on an ATP question and a vesicle question. The other group was assigned a motor protein question and an osmosis question. All students answered final exam questions on all four topics. The Topic Analysis group for each final exam question consisted of all students who completed the corresponding topic-matched homework question (Figure 1). The Control Analysis group for each final exam question consisted of students whose homework questions were not topic-matched to the final exam question. The percentage of students who correctly answered the final exam questions (%Correct on Final) was significantly higher in the Topic Analysis versus the Control Analysis group for two of the four questions (***$p < 0.001$, Fisher's exact test; Figure 1, Discussion Training).

In the 2009 class, there were two training sessions, both conducted in lecture (Lecture Training): one in week 2, after the first in-class quiz, and one in week 5, right after the midterm. Half of the students in each lecture were assigned an analysis homework activity on an osmosis question and a Na^+/K^+ pump question. The other half of each lecture group was assigned a buffer question and a transporter question. Since the groups spanned two lectures with different midterm and final exams, there were two question pairs for each topic (eight original pairs, but one was dropped after final exam grading, as the wording of the question was ambiguous). The percentage of students who correctly answered the final exam questions (%Correct on Final) was significantly higher in the Topic Analysis versus the Control Analysis group for three of the seven questions (* $p < 0.05$, *** $p < 0.001$, Fisher's exact test; Figure 1, Lecture Training).

The significantly higher performance of the Topic Analysis versus Control Analysis group for five of the 11 final exam questions demonstrates a benefit to the class as a whole, since this analysis included all students, regardless of whether they answered the midterm question correctly or not. In addition, in no case was the performance of the Topic Analysis group

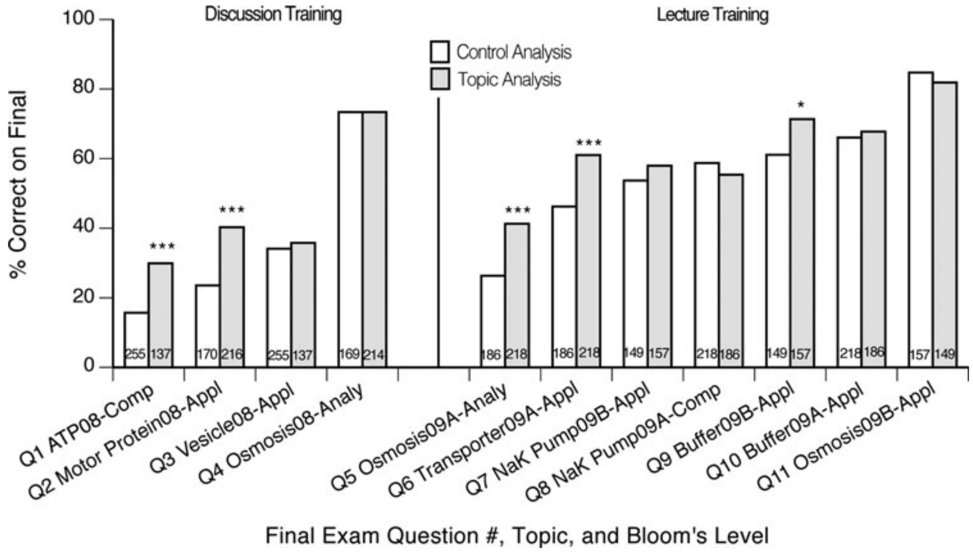


Figure 1. Analysis of assigned midterm questions significantly increases performance on some topic-matched final exam questions. In the Discussion Training in 2008, the percentage of students who answered correctly on the final exam was significantly higher in the Topic Analysis group compared with the Control Analysis group for two of four questions (*** $p < 0.001$, Fisher's exact test). In the Lecture Training in 2009, the percentage of students who correctly answered the final exam questions was significantly higher in the Topic Analysis group compared with Control Analysis group for three of seven questions (* $p < 0.05$, ***, $p < 0.001$, Fisher's exact test). There was no significant difference for the other questions between Topic Analysis and Control Analysis groups. % Correct on Final represents the number of students who answered the final exam question correctly divided by the total number of students in each analysis group (*n*: indicated by numbers in bars, includes students who answered midterm question correctly and incorrectly). Topic Analysis is defined as the group of students who completed an analysis homework on the topic matched to the final exam question. Control Analysis is defined as the group of students who completed an analysis homework on topics related to the other study-related final exam questions. Questions are identified by number, topic (ATP, Osmosis), year of exam (08 or 09), class lecture section (A or B, where relevant), and Bloom's level.

significantly lower than the Control Analysis group. Of the five questions for which the performance of the Topic Analysis group was higher, three were application level, with one each at the level of comprehension and analysis on Bloom's scale. This indicates that the exam analysis exercise increased the ability of students to answer questions at higher levels of cognition.

Student-Choice Analysis Questions

To evaluate the effect of allowing students to select the analysis question, students in 2009 were also instructed to choose one question they missed on the midterm from a list of four possible questions. Therefore, only students who missed the original question were included in this analysis. For each topic, there was one question from lecture A and one question from lecture B. The Topic Analysis students completed a homework on the topic of the final exam question indicated on the graph (Figure 2). The Control Analysis students completed a homework on one of the other three final exam question topics. The percent correct on the final exam question was significantly larger in the Topic Analysis group compared with the Control Analysis group for two out of eight questions, both of which were application-level questions on cell signaling ($p < 0.001$, Fisher's exact test; Figure 2).

From these data, it is clear that the analysis homework on both assigned and student-choice topics resulted in significant improvement for some, but not all, topic-matched questions on the final exam. Therefore, we were interested in identifying factors that impact the effectiveness of the activity.

Similarity of Midterm–Final Exam Question Pairs

Topic-matched question pairs were divided into four categories based on similarity: Identical, Same Emphasis, Moderately Different Emphasis, and Different Emphasis (described in *Materials and Methods*). To determine whether there was a relationship between the degree of similarity within each midterm–final exam question pair and the effectiveness of the activity, the difference in the percent correct on the final exam question between the Topic Analysis and Control Analysis groups was plotted for all questions grouped by similarity category (Figure 3).

To include both the assigned and student-choice questions in this analysis, the data included only students who originally missed the midterm exam question. This reduced the sample size for the assigned questions, and two no longer showed statistically significant differences: Buffer 09B (Q9) and Osmosis 09A (Q5). However, the remaining questions that still showed significant differences between Topic Analysis and Control Analysis groups in this analysis were all in the Same Emphasis category (Figure 3). These data suggest that when the emphasis was the same, even when scientific context and/or wording was changed, the exercise increased the ability of our introductory students to answer a subsequent question on the same topic ~50% of the time (five of nine questions). The difference between the Topic Analysis and Control Analysis groups ranged from 8 to 25% for the five questions, and four of the five questions were at the Bloom's application level of cognition. Since these require the ability to predict outcomes in new situations or interpretation of new data sets, the learning gains are associated with this level of processing, rather than just knowledge-level gains.

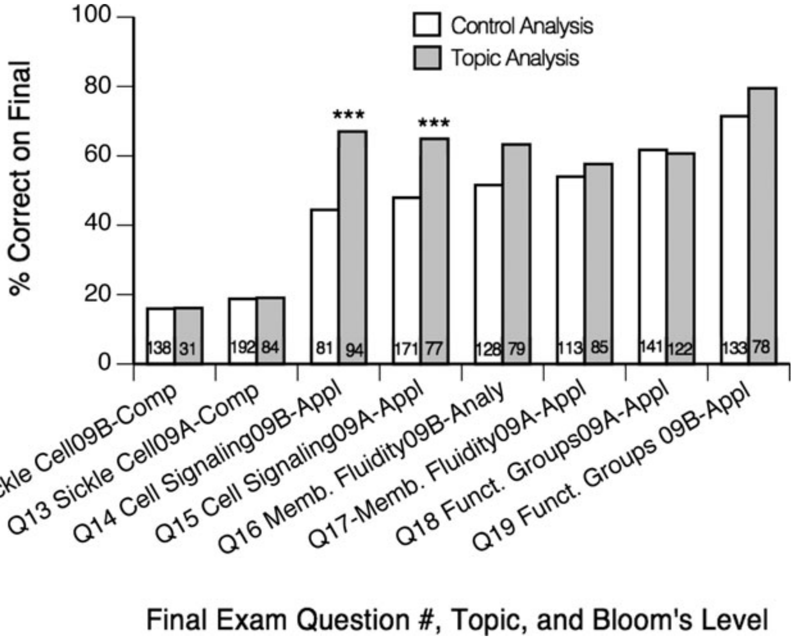


Figure 2. Student choice of an analysis question is not more effective than assignment of analysis questions. The percentage of students who correctly answered the final exam questions was significantly higher for two of eight questions in the Topic Analysis vs. Control Analysis group (*** $p < 0.001$, Fisher's exact test). There was no significant difference between Topic Analysis and Control Analysis groups for the other six questions. Percent correct represents the number of students who answered each question correctly divided by the total number of students in each analysis group. Students in this comparison all initially missed the question on the midterm (*n*: numbers in each bar).

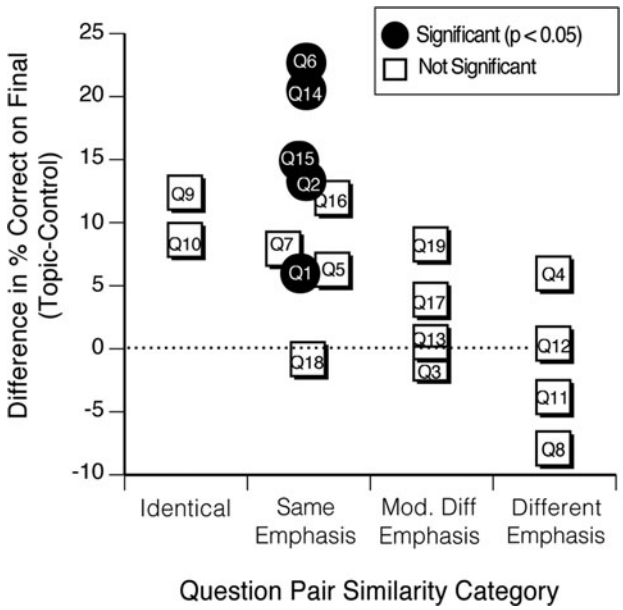


Figure 3. Question-pair similarity is correlated with effectiveness of exam analysis homework. The difference in percent correct between Topic Analysis and Control Analysis groups is plotted for each final question grouped by similarity category of the midterm–final exam question pair. Only students who originally missed the topic question are included in this comparison. In the Same Emphasis category, five of the nine analysis assignments resulted in significantly higher performance of the Topic Analysis group vs. the Control Analysis group on the final exam question (black circles). None of the analysis assignments in the other three categories resulted in significant differences in performance between Topic Analysis and Control Analysis groups. For each individual question pair, significance was evaluated by comparing the number of students who answered correctly in the Topic Analysis group vs. the Control Analysis group on the final exam question ($p < 0.05$, Fisher's exact test). Each point has a designated question number, which refers to the questions identified in Figures 1 and 2.

Neither the two Identical pairs nor the eight pairs classified as Moderately Different Emphasis or Different Emphasis showed a significant effect from the analysis homework. Further evaluation showed that three of the eight Different Emphasis pairs, and both of the Identical pairs, had final exam questions that were relatively easy based on the percentage of the Control Analysis group answering the question correctly (60% or more of the Control Analysis group answered final exam question correctly; see Figures 1 and 2). One of the four Same Emphasis questions that did not show a significant effect of the analysis was difficult (Q5), and the rest had 50–60% of Control Analysis students answering correctly (Q7, Q16, Q18). These data indicate that the exercise is not effective if the final exam question emphasizes a different aspect of the topic from the homework question. Additional data will be required to determine whether there is a significant correlation between question difficulty and exercise effectiveness.

Quality of Student Midterm Question Analyses

We also explored the potential role of the quality of the completed analysis on the effectiveness of the activity by classifying each homework response as strong or weak (see rubric in *Materials and Methods*). To determine whether quality of student analysis is a good predictor of increased ability to answer the topic-matched final exam question, we limited our comparisons with students who missed the homework analysis question on the midterm.

For the assigned topics, the percentage of students within the Topic Analysis group who correctly answered the matched final exam question was significantly higher for those turning in a strong versus a weak analysis for five of 11 questions (* $p < 0.05$, ** $p < 0.01$, ***$p < 0.001$, Fisher's exact test; Figure 4). In no case was the performance of the strong analysis group significantly lower than the weak analysis group. These data indicate that the quality of the homework analysis is also correlated with activity effectiveness.

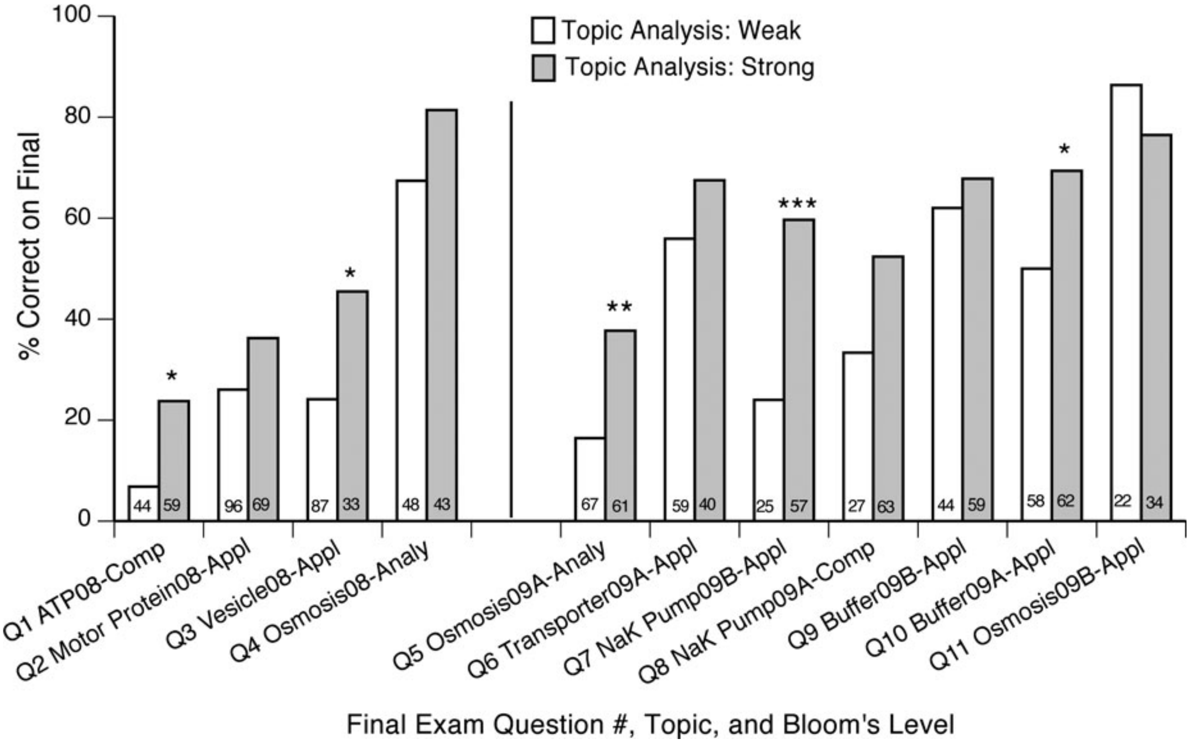


Figure 4. Quality of completed homework is correlated with effectiveness of the assignment. Within a Topic Analysis group, students submitting a homework classified as strong performed significantly better than students turning in a homework classified as weak on five of 11 assigned topics. Percent correct represents the number of students who answered each question correctly divided by the total number of students in that analysis group, all of whom initially missed the question on the midterm (indicated by numbers in each bar). (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, Fisher's exact test).

For the student-choice topics, it was not possible to complete an analysis-quality comparison, because the number of students who turned in weak analyses for the questions was too small for a statistically meaningful Fisher's exact test (questions averaged 15% for weak responses). This was in contrast to the assigned analysis, where the number of strong and weak analyses was similar for each topic (averaging 43% weak). This suggests that students will elect to analyze questions they believe they can answer correctly when given the choice, rather than the question they need the most help understanding.

Student Attitude

The effectiveness of an activity can also be influenced by student attitude. A postclass survey on teaching techniques garnered 446 anonymous responses from a total of 796 students in 2009. Of the respondents, 76% agreed the LFE activity was helpful, and 74% agreed the activity made them examine their graded midterm more carefully (Table 3). More than half of the students (54%) felt that completing their own analyses taught them more biology than reading the instructor-annotated answer key (provided after homework was submitted), and 78% felt they had sufficient information from notes and readings to construct good analyses. In addition, more than half of the students (51%) indicated they will use this technique on their own in future classes, suggesting that the value of using this strategy to increase learning gains extends beyond the class in which it is introduced (Table 3).

DISCUSSION

The availability of formative assessment tools that allow students to evaluate and adjust their performance while learning is often limited in large-enrollment biology classes. However, most classes give summative midterms, and many give cumulative final exams. In this study, we describe an independent midterm exam analysis activity that is easy to implement, and we show it can, in some cases, improve student performance on final exam questions matched by topic to the homework. We discuss in the following sections the benefits of the experimental design, factors that influenced the success of the analysis activity, and how our results will guide development of new strategies to increase learning in our large classrooms.

Assessment with Topic-Matched Midterm–Final Exam Question Pairs

Two studies assessing the effectiveness of midterm analysis activities on learning documented increases in student performance based on retakes of the original exam (Wininger, 2005; Drouin, 2010). One caveat noted was the difficulty in determining whether students who performed better on the retest had gained a deeper understanding of the material or whether they recognized the correct answers to previously seen questions (Drouin, 2010). To address this issue, we employed an assessment strategy that evaluated performance on final exam questions that were matched by topic to

Table 3. Student survey responses to exam question analysis activity

	Strongly agree[a]	Agree[a]	No opinion[a]	Disagree[a]	Strongly disagree[a]	Number of student responses
Completing the LFE activities helped me learn the course material.	20	56	14	10	0	441
The LFE assignment made me examine my graded midterm exam more carefully.	25	49	14	11	1	437
I had sufficient information available from my notes and the reading to formulate good biological explanations on the LFE assignments.	20	58	16	6	0	437
I learned more biology from completing the LFE assignment on a question I missed compared to reading the annotated key for a question I missed.	13	41	26	18	3	437
I plan to study from future graded exams "LFE-style" by writing out why the wrong answers were wrong and the right answers were right.	14	37	32	15	2	439

[a] All values are given in percentages.

homework questions from the midterm, but the questions could vary in phrasing, scientific context, or topic emphasis.

There are several advantages to this experimental design. First, it controlled for the possibility of students simply recognizing answers to questions they had previously seen. Second, this study was conducted in an ecologically relevant setting with high-stakes summative exams in which the motivation to answer correctly is both high and consistent. In addition, because the homework exercise was completed in week 6 and the assessment was in week 11, the learning gains observed are relatively long term. Finally, the exercise focused on questions that students had the most difficulty with on the midterm exam, and these were predominantly application- and analysis-level questions. This is consistent with our introductory students having reasonable facility with learning new knowledge with some integrative understanding but having more difficulty with the higher levels of cognition. The exercise aligns the faculty desire for students to increase their mastery of the course material at the application and analysis level with the students' desire to increase their grades, and is ecologically valid, as the activity/assessment is embedded in the normal curriculum (Black and Wiliam, 1998).

Within-Section Control

Education research conducted in the classroom often compares the effect of a new teaching strategy or activity on the performance of students in one class with the performance of students in another class, taught either concurrently or at different times by the same instructor(s) (Allain and Williams 2006; Armstrong *et al.*, 2007; Freeman *et al.*, 2007; Moravec *et al.*, 2010). When the instructors are the same and the demographics of the comparison classes are similar, these comparisons can provide important insights into the changes in performance associated with implementation of the new teaching strategy. Although we conducted our experiment in three different sections of our introductory lecture class, our within-class controls allowed us to assess the effectiveness of a new technique in a single section. To ensure that students within a single section were not given an unfair advantage by being assigned to an experimental versus a control group, each student was a member of the experimental group for one subset of questions and a member of the control group for the other subset of questions. Increases in the final exam performance on specific questions could therefore be attributed to the analysis activity and not to inherent differences between the student groups.

A Comment on "the Testing Effect"

The act of retrieving an answer from memory, as occurs when students are tested on the same or even similar material multiple times, can increase learning independently of other interventions (Karpicke and Roediger, 2008; Karpicke and Blunt, 2011). Our homework, however, does not increase learning due to a testing effect, because the assignment is not done from memory. All members of both groups were tested on each topic twice, once on the midterm and once on the final.

Novice Learners and Question Similarity

Like many faculty, we return all exams and provide students access to exams from previous years to help them prepare for upcoming tests. This requires writing new exams on the same set of core concepts each time the class is taught. The question variants in this study were chosen to represent the range we have routinely used when writing new questions on the same topic, unlike previous work that has focused on pre- and postactivity performance on identical questions (e.g., Wininger [2005]; Drouin [2010]). After completing this experiment, we found it illuminating that although the questions were carefully topic matched, students were easily confused by shifts of emphasis within a topic.

Of the topic-matched questions that had the same emphasis as the homework, approximately 50% of them were associated with significant increases in student performance. Notably, the Bloom's levels of these questions were primarily at the application level, indicating that our students can, in some cases, apply knowledge/problem-solving skills (including data interpretation) learned in one setting to a different context. However, when the emphasis in the topic-matched questions was different, performance did not increase significantly.

These results are consistent with previous studies showing that novices rely more on surface features for identifying categories of questions, while experts recognize deeper features, such as underlying principles or concepts (Chi *et al.*, 1981; Cummins 1992). Because our introductory biology students are novice learners, they are likely to be relying more heavily on surface features, which makes it difficult for them to transfer information learned in one context to another application (Barnett and Ceci, 2002; Wagner, 2006). Distinctions between surface and deep features may be especially problematic in biological sciences, because core topics are broad and multifaceted. For example, for the topic of cell membrane fluidity, there are multiple determinants influencing the degree of fluidity, including lipid composition, protein content, and temperature. Membrane fluidity can in turn affect distinctly different cell processes, such as protein movement within a membrane and cell movement from place to place. Thus, it is important for us to teach students study techniques that facilitate a broad understanding of core topics.

Teaching novice learners to recognize underlying concepts for problem solving requires careful guidance of the learners (Novick, 1988; Kirschner *et al.*, 2006). A recent study involving analysis of mistakes on physics quiz questions suggests that having students complete a self-diagnosis using a solution outline and a detailed diagnosis rubric can improve learning outcomes (Yerushalmi *et al.*, 2008). As we develop our exam analysis tool further, we plan to provide more scaffolding to the students in the independent analysis assignment to determine whether the effectiveness can be improved. Sorting problems into categories and making analogical comparisons are approaches that have also been used to help students recognize underlying concepts (Chi *et al.*, 1981; Gick and Holyoak, 1983; Quilici and Mayer, 2002). In the future, we plan to evaluate the effectiveness of having students categorize exam problems into core topic groups. This could be done by providing groups of students with envelopes containing exam questions printed on individual note cards, with instructions to sort the questions by topic before answering them. If categorization of questions into topic groups provides students with meaningful insights into the scope of information and logic required to master each topic, this should not only lead to improved exam performance, but it will help introductory students begin developing a broad conceptual framework of biology early in their education.

Importance of Quality of Submitted Analyses

Comparison within a Topic Analysis group showed that students turning in strong analyses were significantly more likely to answer the final exam question correctly than those turning in weak analyses for a number of the assigned topics. This indicates that the quality of the work completed, and not just an additional exposure to a topic, can affect performance on the final exam. While we do not have information about what influenced homework quality, 78% of the students in a postclass survey indicated they had sufficient information available from their notes/reading to formulate good biological explanations on the analysis assignments. Interestingly, the quality of the homework submitted for assigned questions was classified as strong only 57% of the time. This suggests that approximately 20% of our students are unable to recognize a weakness in their understanding of a biological concept and have difficulty in independently acquiring the appropriate information needed to increase their understanding. The disconnect between perception and performance suggests that these students need additional guidance in recognizing weaknesses in their understanding, a skill important for successful self-regulated learning (Butler and Winne, 1995; Ley and Young, 2001).

When students were allowed to choose a question to analyze, the strong responses increased to 85%, indicating the students were selecting questions they were most likely to answer correctly. It is not clear what was driving this choice, but possible explanations include students gravitating toward questions for which they had more background knowledge or for which information was more readily available. Nevertheless, based on the prevalence of strong analyses for the student-choice homework questions, one would expect a larger percentage of these homework questions to result in significant difference between the Topic Analysis and Control Analysis groups on the matched final exam questions than we observed (Q14 and Q15, Figure 2). This discrepancy is most likely explained by the fact that only four of the eight student-choice questions had the characteristics of final exam questions associated with high efficacy: same topic emphasis as homework problem and relatively difficult (Q14, Q15, Q16, Q18). The homework exercise was effective for the two most difficult questions of the four; while the number of questions is small, this is consistent with strong analyses increasing the probability of the activity resulting in significant differences between the Topic Analysis and Control Analysis groups.

Self-Regulated Learning

In the face of rapid scientific discovery and constantly evolving technology, it is critical that our biology students develop into effective, self-regulated learners. This requires that students be able to monitor, manipulate, and improve their own learning using strategies such as resource management, goal setting, positive beliefs, and self-reward (Schunk, 1990; Ley and Young, 2001). Formative assessment opportunities are important in facilitating self-regulated learning (Butler and Winne, 1995; Nicol and MacFarlane-Dick, 2006). The exercise we report was simple and low cost and resulted in increased performance on authentic, high-stakes graded elements in the class. Therefore, it can be used in very large classes, where formative assessment opportunities are often limited. Just over half of the students in our class indicated they would apply the exam analysis strategy described in this study in future classes. Thus, helping students employ this technique effectively has the potential to increase learning gains beyond the class in which it is introduced.

ACKNOWLEDGMENTS

We thank Roger Roxas for his excellent help in scoring the student analyses, all of our students in Bio 93 2005–2009, and Dr. Rahul Warrior for his support during his instruction of Bio 93. A special thanks as well to UCI's Distribution and Document Management Services and Norm Franz and Linh Nguyen for the development of the electronic exam return service Rapid Return. This research was funded by a grant from the Howard Hughes Medical Institute Professor Program to D.K.O.D.

REFERENCES

Allain R, Williams T (2006). The effectiveness of online homework in an introductory science class. J Coll Sci Teach *35*, 28–30.

Allen D, Tanner K (2005). Infusing active learning into the large-enrollment biology class: seven strategies, from the simple to complex. Cell Biol Educ *4*, 262–268.

Armstrong N, Chang S-M, Brickman M (2007). Cooperative learning in industrial-sized biology classes. CBE Life Sci Educ *6*, 163–171.

Barnett SM, Ceci SJ (2002). When and where do we apply what we learn? A taxonomy for far transfer. Psychol Bull *128*, 612–637.

Black P, Wiliam D (1998). Assessment and classroom learning. Assess Educ Princ Pol Pract *5*, 7–74.

Bloom BS, Engelhart MD, Furst EJ, Hill WH (1956). The Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook I: Cognitive Domain, New York: McKay.

Boehm R, Gland JL (1991). Using exams to teach chemistry more effectively. J Chem Educ *68*, 455.

Bolt-Lee C, Foster SD (2000). Examination retakes in accounting: increasing learning by writing after the exam. LLAD *4*, 40–46.

Boud D (2000). Sustainable assessment: rethinking assessment for the learning society. Stud Contin Educ *22*, 151–167.

Butler DL, Winne PH (1995). Feedback and self-regulated learning: a theoretical synthesis. Rev Educ Res *65*, 245–281.

Caldwell JE (2007). Clickers in the large classroom: current research and best-practice tips. CBE Life Sci Educ *6*, 9–20.

Chi MTH, Feltovich PJ, Glaser R (1981). Categorization and representation of physics problems by experts and novices. Cogn Sci *5*, 121–152.

Cronbach LJ (1951). Coefficient alpha and the internal structure of tests. Psychometrika *16*, 297–334.

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

Cummins DD (1992). Role of analogical reasoning in the induction of problem categories. Learn Mem *18*, 1103–1124.

De Champlain AF (2010). A primer on classical test theory and item response theory for assessments in medical education. Med Educ *44*, 109–117.

Ding L, Chabay R, Sherwood B, Beichner R (2006). Evaluating an electricity and magnetism assessment tool: brief electricity and magnetism assessment. Phys Rev ST Phys Educ Res *2*, 010105.

Drouin MA (2010). Group-based formative summative assessment relates to improved student performance and satisfaction. Teach Psychol *37*, 114–118.

Freelon DG (2010). Recal: intercoder reliability calculation as a web service. Int J Internet Sci *5*, 20–33.

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.

Gick ML, Holyoak KJ (1983). Schema induction and analogical transfer. Cogn Psychol *15*, 1–38.

Hoskins SG, Stevens LM (2009). Learning our L.I.M.I.T.S.: less is more in teaching science. Adv Physiol Educ *33*, 17–20.

Karpicke JD, Blunt JR (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. Science *331*, 772.

Karpicke JD, Roediger HL (2008). The critical importance of retrieval for learning. Science *319*, 966–968.

Kirschner PA, Sweller J, Clark RE (2006). Why minimal guidance during instruction does not work: an analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. Educ Psychol *41*, 75–86.

Kitchen E, Bell JD, Reeve S, Sudweeks RR, Bradshaw WS (2003). Teaching cell biology in the large-enrollment classroom: methods to promote analytical thinking and assessment of their effectiveness. Cell Biol Educ *2*, 180–194.

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

Ley K, Young DB (2001). Instructional principles for self-regulation. Educ Tech Res Devel *49*, 93–103.

Moravec M, Williams A, Aguilar-Roca N, O'Dowd DK (2010). Learn before lecture: a strategy that improves learning outcomes in a large introductory biology class. CBE Life Sci Educ *9*, 473–481.

Nicol DJ, Macfarlane-Dick D (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. Stud High Educ *31*, 199–218.

Novick LR (1988). Analogical transfer, problem similarity, and expertise. J Exp Psychol Learn Mem Cogn *14*, 510–520.

Quilici JL, Mayer RE (2002). Teaching students to recognize structural similarities between statistics word problems. Appl Cogn Psychol *16*, 325–342.

Sadler DR (1989). Formative assessment and the design of instructional systems. Instr Sci *18*, 119–144.

Schunk D (1990). Goal setting and self-efficacy during self-regulated learning. Educ Psychol *25*, 71–86.

Wagner JF (2006). Transfer in pieces. Cogn Instr *24*, 1–71.

Wininger SR (2005). Using your tests to teach: formative summative assessment. Teach Psychol *32*, 164–166.

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

Yerushalmi E, Mason A, Cohen E, Singh C, Henderson C, Sabella M, Hsu L (2008). Effect of self diagnosis on subsequent problem solving performance. ed. C Henderson, M Sabella, and L Hsu, 2008 Physics Education Research Conference Proceedings, AIP Conference Proceedings, vol. 1064, 53–56.