CBE—Life Sciences Education
Vol. 6, 132–139, Summer 2007

Article

Prescribed Active Learning Increases Performance in Introductory Biology

Scott Freeman, Eileen O'Connor, John W. Parks, Matthew Cunningham, David Hurley, David Haak, Clarissa Dirks, and Mary Pat Wenderoth

Department of Biology, University of Washington, Seattle, WA 98195

Submitted September 25, 2006; Revised February 12, 2007; Accepted February 13, 2007
Monitoring Editor: Martha Grossel

We tested five course designs that varied in the structure of daily and weekly active-learning exercises in an attempt to lower the traditionally high failure rate in a gateway course for biology majors. Students were given daily multiple-choice questions and answered with electronic response devices (clickers) or cards. Card responses were ungraded; clicker responses were graded for right/wrong answers or participation. Weekly practice exams were done as an individual or as part of a study group. Compared with previous versions of the same course taught by the same instructor, students in the new course designs performed better: There were significantly lower failure rates, higher total exam points, and higher scores on an identical midterm. Attendance was higher in the clicker versus cards section; attendance and course grade were positively correlated. Students did better on clicker questions if they were graded for right/wrong answers versus participation, although this improvement did not translate into increased scores on exams. In this course, achievement increases when students get regular practice via prescribed (graded) active-learning exercises.

INTRODUCTION

Recent efforts to improve science education at U.S. universities have focused on three themes: 1) faculty should apply the same hypothesis-testing framework in their teaching as they do in their benchwork and fieldwork (Handelsman *et al.*, 2004; Cech and Kennedy 2005); 2) the overall number of undergraduate and graduate degrees awarded to U.S. citizens in science, mathematics, and engineering needs to increase dramatically (e.g., Cech and Kennedy, 2005); and 3) more underrepresented minorities (URMs) and students from disadvantaged backgrounds need to be recruited and retained in science and technology majors (Matsui *et al.*, 2003; Summers and Hrabowski, 2006). Our work was inspired by these themes and addressed a simple question: Can the high failure rate in an introductory biology course for majors be reduced?

Our approach grew out of the first author's experience while teaching the initial quarter of a year-long sequence for University of Washington (UW) students who intend to major in biology and/or apply to health-related professional schools. In spring 2002, the instructor taught the course in a modified Socratic style, stopping the lecture frequently to ask questions and not proceeding until one or more students had responded. Although student ratings of the course and instructor were high, 19.6% of the students did not qualify to proceed in biology; almost half failed to do well enough to declare the major. In response, the instructor introduced an array of daily active-learning, in-class exercises in spring 2003. These tasks included case history problems done by informal groups, think/pair/share exercises (Lyman, 1981), muddiest point writing (Mosteller, 1989; Angelo and Cross, 1993; Boyd, 2001), exam-question writing, minute papers (Angelo and Cross, 1993), and discussions of exam questions from previous quarters. These exercises were not graded. Although the course still received high ratings, the failure rates did not show significant improvement.

High failure rates in "gateway" courses are unacceptable for two reasons: they contribute to low graduation rates and extended time-to-graduation for the institution as a whole, and they have a disproportionately large impact on URMs and other students from disadvantaged backgrounds. A review of students enrolled in Biology 180 during a recent

DOI: 10.1187/cbe.06–09–0194
Address correspondence to: Scott Freeman (srf991@u.washington.edu).

3-yr period (2001–2003) showed that on average, approximately 40% of URMs and economically disadvantaged students received a grade below 2.0 or withdrew before completing the course (Dirks and Cunningham, 2006). URM students often come from high schools that have not been as demanding as those attended by white or Asian students (Cota-Robles and Gordan, 1999; Gandara and Maxwell-Jolly, 1999). Teaching techniques that increase achievement by the most poorly prepared students should be an effective way to increase recruitment and retention of minorities in the natural sciences.

On the basis of our own observations and extensive interviews with students, we generated the following explanations for the traditionally high failure rate in this course:

- ESL (English as a second language) and other students struggle with written exams;
- Most students were being asked to answer exam questions written at the application and analysis levels of Bloom's Taxonomy of Learning Objectives (Bloom, 1956) for the first time;
- Many students do not understand the time commitment required to succeed in the course; and
- Students learn better if they are active, but most prefer being passive.

To test these ideas, we designed, implemented, and evaluated a series of course designs that attempted to accomplish two goals: 1) offer extensive practice with written and multiple-choice questions above the bottom rung of Bloom's taxonomy; and 2) prescribe active participation by grading weekly and daily practice. Our fundamental premise was that active learning increases performance on exams because it gives students opportunities to practice and that introductory students must be required to practice by awarding points.

MATERIALS AND METHODS

The course in question enrolls ~345 students and is usually taken by sophomores. A typical demographic makeup is 58% female, 46% white, 30% ESL, and 7% URM students, of whom ~17% are in the Educational Opportunity Program (EOP). EOP students have been identified by the UW as economically or educationally disadvantaged. The course is offered every quarter of the year and is taken by a total of ~1200 students per year or almost 25% of an average incoming class at our university. The course prerequisites are two quarters of inorganic chemistry.

On a grading scale from 0.0 to 4.0, a grade of 1.5 or higher in the initial course is required to register for the next course in the year-long (3-quarter) sequence. Thus, students who do not receive a grade of 1.5 or higher fail to advance in the life sciences. Students who do not receive 2.5 or higher must average 2.0 or higher over the three courses in the series in order to declare biology as their major. Thus, students who receive less than 2.5 are at high risk of failing to advance in biology or other life sciences majors.

The study was organized in three steps: 1) analyzing the characteristics of students who had taken the course previously to better understand the reasons for failure and to predict student performance a priori; 2) implementing four contrasting course designs during the spring quarter 2005; and 3) repeating one of these four course designs and contrasting it with a fifth course design in the fall quarter 2005.

Risk Analysis

To better understand the reasons for the high failure rate, we analyzed data on 3338 students who had started the introductory biology series at the UW between the autumn quarters of 2001 and 2004. Specifically, we attempted to correlate the following demographic variables and measures of prior academic performance with failure in one or more of the courses in the series: gender, ethnicity as Caucasian, Asian, or URM (African-American, Native American, Hispanic, or Pacific Islander), chronological age, average grade in UW chemistry classes at the time of entering the course, overall UW grade point average (GPA) at the time of entering the course, UW class standing (freshman, sophomore, etc.) at the time of entering the course, high school GPA, Scholastic Aptitude Test (SAT) verbal score, SAT math score, American College Test (ACT) score, score on the UW math placement test given to matriculating students, Test of English as a Foreign Language (TOEFL) score, and EOP status.

After analyzing these variables for covariation and evaluating them for missing data, we performed a factor analysis to determine which variables could be omitted from the analysis or aggregated into a single index. These steps led us to drop UW chemistry GPA, high school GPA, ACT score, TOEFL score, and UW math placement score from the model.

To determine which of the remaining variables were most capable of predicting failure in the course, we performed bivariate logistic regression with backward elimination. After determining which variables were most important, we used them to design a regression model that predicted student grades in the first course in the sequence. In this way, we could identify students who were at low risk or high risk of failing the course. We defined high-risk students as those predicted to get below a grade of 2.5 in the course, and low-risk students as those predicted to get a grade of 2.5 or higher.

Spring 2005 Course Design

The spring 2005 course was listed as two equal-size sections during registration, so that students signed up for sections unaware of contrasting course designs. The two sections were taught back-to-back by the same instructor (S.F.) in the same room, using identical notes. Students from the two sections were mixed randomly in lab sections and required field trips.

During each class the instructor posed four multiple-choice questions that required a response from all students. The period started with a question based on the previous session's material and a question on the reading for that day. Twice during a lecture delivered in a modified Socratic style, the instructor posed questions based on the material being discussed and that had to be answered by all students. The questions were designed to be difficult. Specifically, they attempted to either test students' ability to analyze an aspect of the topic or apply a concept in a new situation. If less than ~60% of the answers to these questions were correct, the instructor told the students to discuss the question among themselves and then reanswer (this is the peer-instruction technique; see Mazur, 1997; Crouch and Mazur, 2001). Students answered from four to eight formal questions each day, with an average of 5.6.

In the "clickers section," students were given an electronic response device that they registered with their name and student number. After each class session, the instructor would choose three of the four to eight responses to grade. Correct answers on these three questions were worth 1 point each. A total of 100 clicker points were possible for the quarter, representing ~14% of the total points.

In the "cards section," students were given four cards with A, B, C, or D printed on them. Students held up these cards to answer the same in-class questions posed to the clicker section. Because other students could look at the cards if they desired, card responses were public. To enforce or prescribe participation, the instructor "stared down" students who occasionally did not hold up a card, so that virtually all students responded to all questions. Card responses were public but ungraded; clicker responses were private but graded.

The full class met 4 d per week; during the fifth class period each week students were given five written, exam-style questions to complete in 35 min. Answers were then randomly assigned to another student for a 15-min grading period based on a key and rubric provided by the instructor. Correct answers were to be given 2 points, partial credit answers 1 point, and blank or unintelligible answers 0 points. Both the answerer and grader were anonymous to each other; only course staff knew their identities. Students did nine of these practice exams for a total of 90 possible points—roughly 15% of the total grade.

Students in the cards and clickers sections were randomly assigned to one of two methods of completing identical practice exams. Within each section, half the students did the practice exams and grading by themselves online, and half did the practice exams and grading as part of a study group. Students who did the assignment online could do so anywhere, but had to log in and submit answers and grades on the 35-min + 15-min schedule during the class period. Students assigned to study groups met in a classroom on campus and were proctored by a staff member. The staff member did not answer content questions or assist the groups in any way. The staff member distributed the hard-copy questions, accepted answers after 35 min, randomly assigned answer sheets for the 15-min grading period, and collected the graded sheets.

Students were assigned to study groups by the instructor, on the basis of their course grade predicted by the regression model from the risk analysis. Each study group had one student who was predicted to receive below 1.5 in the course, two students who were predicted to receive between 1.5 and 3.0, and one student who was predicted to receive 3.0 or higher. Students were unaware of this structure, however. Each week, study group members signed up to serve as their group's "manager," who coordinated the exercise; "strategist," who considered ways to approach each question; "recorder," who wrote the answers; or "encourager," who gave positive feedback to participants. These roles were explained by the staff proctor but were not enforced by peer evaluation or other techniques.

To summarize, the spring quarter tested the following four designs: clickers + online practice, clickers + study group practice, cards + online practice, and cards + study group practice. All students took a common final and a common second midterm. On the first midterm, several questions on the exams given to the two sections were identical or formally equivalent. Thus, there were a total of 336 out of 400 total exam points available to use in evaluating student performance on identical exam questions. Students who dropped the course or who were caught cheating on exams were not included in any of the analyses.

We also collected data on attendance. This was done automatically from clicker responses and from the cards section by counting the number of students present during each class period. Although we could only evaluate overall class attendance in the cards section, in the clickers section we could also analyze the number of classes attended by each student.

Fall 2005 Course Design

As in spring 2005, students registered blindly for two equal-sized sections in the fall 2005 course. The sections were again taught back-to-back by the same instructor (S.F.) in the same room using identical notes, with students from the two sections mixed randomly for labs and field trips. In both sections, students completed and graded weekly practice exams by themselves online.

All students in the course were required to purchase a clicker and register it to their name and student number. The instructor posed identical, daily, in-class, multiple-choice questions to each section. In one section, the questions were graded for right/wrong answers using the same format as spring 2005. In the other section, students were given points for participation, with two points per day possible if students answered all questions posed—irrespective of whether their answers were correct or incorrect.

To summarize, the fall quarter repeated the clickers (graded) + online course design and added a clickers (participation) + online course design. Students again took a common final exam; although the sections were given different midterm exams, enough midterm questions were identical or formally equivalent that there were a total of 335 of the 400 total exam points to use in evaluating student performance on common exam questions. Unless otherwise noted, all statistical tests reported here are two-tailed tests.

RESULTS

Risk Analysis

For the first course in the year-long introductory sequence, the logistic regression analysis identified four variables that had a significant effect in predicting the probability of receiving <1.5 in the course when all 3338 students were considered: student age, SAT math, SAT verbal, and UW GPA. Older students were at higher risk of failing; students with higher SAT math and verbal scores and higher UW GPAs were at lower risk of failing. The Wald *t*-statistic coefficients were much higher for UW GPA and SAT verbal than for age and SAT math, however. When we did the same analysis on individual-year student cohorts, we found that UW GPA and SAT verbal were also the most stable year-to-year predictors of student risk of failure. Thus, UW GPA and SAT verbal were the most important variables in predicting student success in the course. On average, students who got below 1.5 in the initial course entered with a UW GPA of 2.60 and an SAT verbal score of 488; students who got above 1.5 in this course entered with a UW GPA of 3.24 and an SAT verbal score of 586.

On the basis of this analysis, we developed a linear regression model to predict the grade that incoming students were likely to receive in the initial course. The model is as follows:

$$\text{Predicted grade} = (0.00291 * [\text{SAT verbal}]) + (1.134 * [\text{UW GPA}]) - 2.663$$

Figure 1 shows the relationship between predicted and actual grades for students in spring 2005 ($R^2 = 0.58$, slope = 1.01, n = 320); the regressions for the fall 2005 sections were similar ($R^2 = 0.60$, slope = 1.18, n = 325). In addition, an

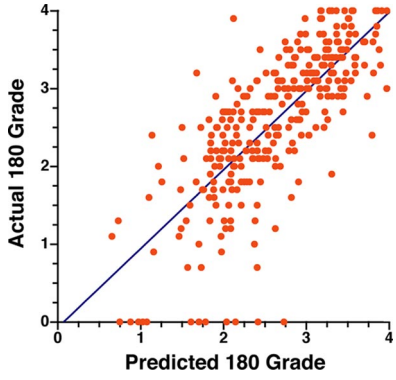

Figure 1. A grade predictor model based on overall UW GPA and SAT verbal score predicts actual grades efficiently ($R^2 = 0.58$, slope = 1.01, n = 320, $p < 0.0001$).

Table 1. Risk of failure by ethnicity and EOP status

	Low risk (%)	High risk (%)
Ethnicity		
Caucasian	58 (297)	36 (130)
Asian	37 (190)	54 (199)
URM	5 (28)	10 (36)
Totals	100 (515)	100 (365)
EOP status		
Non-EOP	91 (531)	69 (274)
EOP	9 (50)	31 (123)
Totals	100 (581)	100 (397)

Values are expressed a percentage, with n in parentheses. Chi-square tests, $p < 0.001$.

ANOVA indicated that the predicted grades for students in the spring 2003, spring 2005, and fall 2005 did not differ significantly (data not shown), indicating that student populations were comparable in these three quarters. Note that these and subsequent analyses do not include students who dropped the course or were caught cheating on exams or assignments; for a small number of students who were admitted to the university without an SAT score, we substituted the class average SAT verbal score to generate a predicted grade.

Because some instructors choose to focus their time and attention on high-achieving students who are most likely to attend graduate or professional school, it is instructive to analyze who the high-risk students are. In our course, 56% of URM students and 71% of EOP students are at high risk of failing (Table 1). URM and EOP students are much more likely than Caucasian, Asian, or non-EOP students to be at high risk of failing (Chi-square test, $p < 0.001$ for both ethnicity and EOP status).

Spring 2005 Course Design

The percentage of students who failed to receive at least 1.5 in the course declined from 15.6% in spring 2003, when the course was taught in a modified Socratic style, to 10.9% in spring 2005, when the course was taught with prescribed active-learning techniques. This drop in failure rate was significant (Fisher's exact test, one-tailed $p = 0.049$).

Exam scores also indicate that as a whole, students in the spring 2005 course did much better in the course than in spring 2003. Out of 400 possible exam points, students in spring 2005 earned on average 14 more points than students in spring 2003 (t test, $p < 0.001$). Although questions from previous exams had not been reused in the past, the spring 2003 and spring 2005 classes were administered an identical midterm exam. The average on this exam in spring 2005 was 11 points higher than in spring 2003 (t test, $p < 0.001$; see Figure 2).

Our next goal was to evaluate whether any of the four course designs increased student achievement more than others. By chance, average grade predictor scores were much higher in the cards section compared with the clickers section (Table 2). To control for this bias between sections, we analyzed high-risk students—defined as those with a predicted grade <2.5—and low-risk students—defined as those with a predicted grade >2.5—separately. We also performed two-way ANCOVAs, with predicted grade as the covariate, to factor out the effect of unlike student populations when comparing performance between students in the two sections.

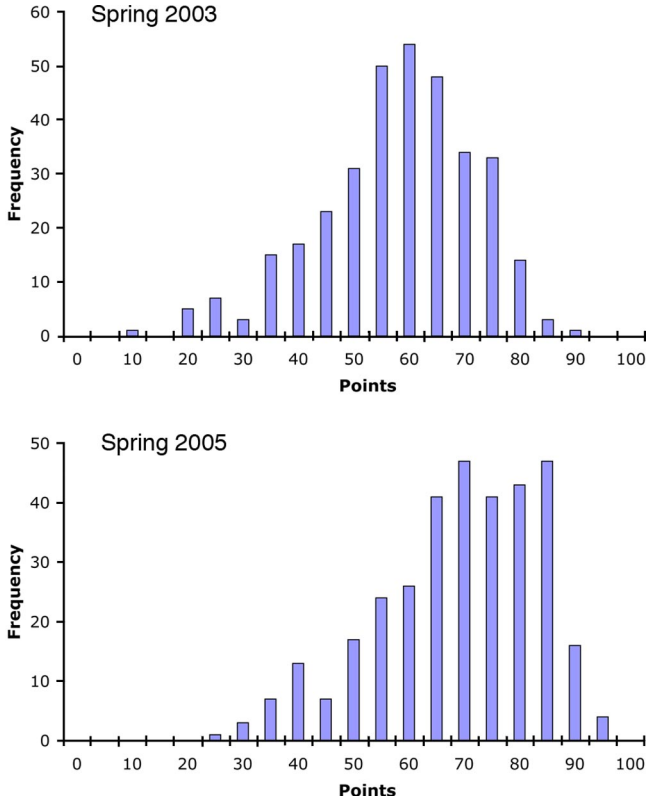

Figure 2. Students in a course (spring 2005) with daily, graded, multiple-choice questions and weekly, graded practice exams did better on an identical midterm exam than students in a course (spring 2003) with daily, ungraded, active-learning exercises (t test, $p < 0.001$).

When we used ANCOVA to evaluate average performance on common exam questions for students in the four course designs (see Table 3), we found no significant effect of practice exam type or in-class question type and no significant interaction between the two variables. Thus, the four course designs worked equally well in terms of boosting student achievement. When we looked at performance on

Table 2. Predicted grades in Biology 180, spring 2005

	Clickers	Cards
Online	2.50 ± 0.65 (87)	2.84 ± 0.74 (80)
Study groups	2.58 ± 0.70 (80)	2.70 ± 0.67 (78)
Totals	2.54 ± 0.67 (167)	2.77 ± 0.71 (158)

Values are mean ± SD, with n in parentheses. Two-way ANOVA, $p = 0.003$.

Table 3. Performance on common exam questions, spring 2005

	All students[a]		High-risk students only[b]		Low-risk students only[c]	
	Clickers	Cards	Clickers	Cards	Clickers	Cards
Online	245 ± 47 (87)	260 ± 48 (80)	223 ± 40 (48)	206 ± 48 (22)	272 ± 41 (39)	281 ± 27 (58)
Study groups	248 ± 45 (80)	253 ± 54 (78)	215 ± 36 (36)	206 ± 51 (31)	275 ± 32 (44)	285 ± 25 (47)
Totals	246 ± 46 (167)	257 ± 51 (158)	219 ± 38 (84)	206 ± 49 (53)	273 ± 36 (83)	283 ± 26 (105)

Values are total points (mean ± SD), with n in parentheses.
[a] Two-way ANCOVA, NS.
[b] ANOVA, $p = 0.08$.
[c] ANOVA, $p = 0.034$.

common exam questions for high-risk students only, the data indicated a weak trend for students in the clickers section to perform better than students in the cards section, although the difference did not reach statistical significance (Table 3; ANOVA, $p = 0.08$). In contrast, low-risk students in the cards section clearly performed better on common exam questions than low-risk students in the clickers section (Table 3; ANOVA, $p = 0.034$). There was no significant difference in performance on common exam questions between students who did practice exams online or in study groups, when either high- or low-risk students were analyzed separately.

The other significant difference between the clickers and cards sections was in attendance. Average attendance was much higher in the clickers section than the cards section (Figure 3A; paired t test, n = 34, $p < 0.0001$). Within the clickers section, attendance had a significant effect on predicting final grade (Figure 4; $R^2 = 0.24$, n = 173, $p < 0.0001$).

Fall 2005 Course Design

In fall 2005, the class average for total exam points (out of 400 possible) was nearly identical to the class average in spring 2005 and significantly higher than spring 2003 (ANOVA, $p < 0.001$). This result replicates the improvement in exam performance shown in the course designs of spring 2005. Because of differences in performance in other parts of the course, however, the percentage of fall 2005 students receiving a course grade <1.5 increased slightly to 11.7%. This failure rate was not significantly different from the percentage in spring 2003 (Fisher's exact test, one-tailed $p = 0.09$).

To compare achievement between the graded and participation sections, we again used ANCOVA with predicted grade as a covariate. This approach allowed us to control for a clear but chance disparity between the predicted grades in the graded and participation sections, with students in the participation section having higher average predicted grades than students in the graded section (Table 4). A one-way ANCOVA showed no difference in common exam points between sections, when all students were considered (Table 5). t tests showed that there was also no difference in performance on common exam questions when only low-risk students or only high-risk students were compared between sections (Table 5). These results suggest that the two course designs worked equally well in improving student performance on exams.

In contrast, there was a clear difference in how students in the two sections performed on clicker questions (Table 6). A one-way ANCOVA, with predicted grade serving as the covariate, showed that students in the graded section did significantly better on clicker questions than students in the participation section ($p < 0.001$). t tests showed that this result held if all students were considered irrespective of the predicted grade differences between sections and when only high-risk students or only low-risk students were compared between sections (Table 6). A paired t test showed that there was no difference between sections in attendance (Figure

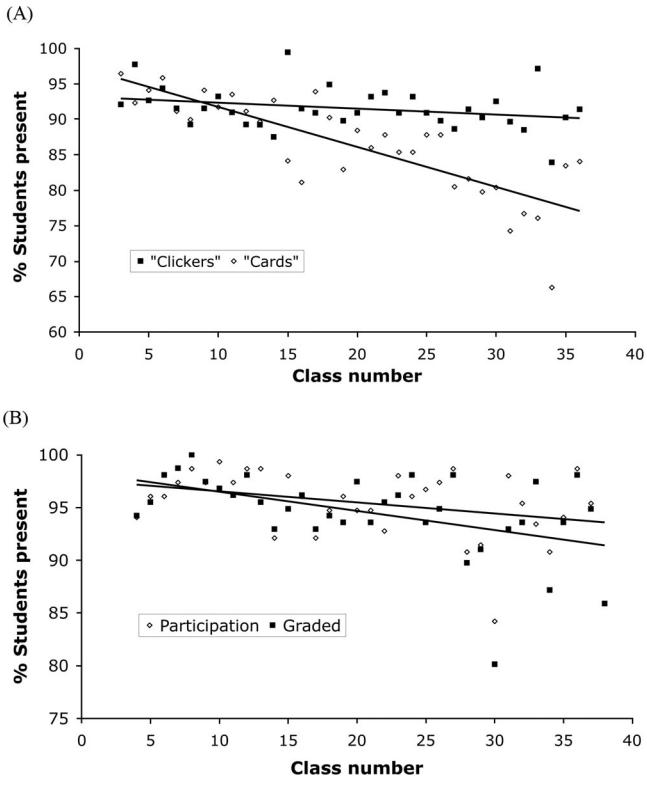

Figure 3. Use of graded in-class exercises increased attendance. (A) Average daily attendance was higher in the section with graded ("clickers") versus ungraded ("cards") in-class questions (paired t test, n = 34, $p < 0.0001$). (B) There was no difference in average daily attendance in classes with in-class questions graded by correct answer versus participation (paired t test, $p > 0.05$). Regression lines are shown in both graphs.

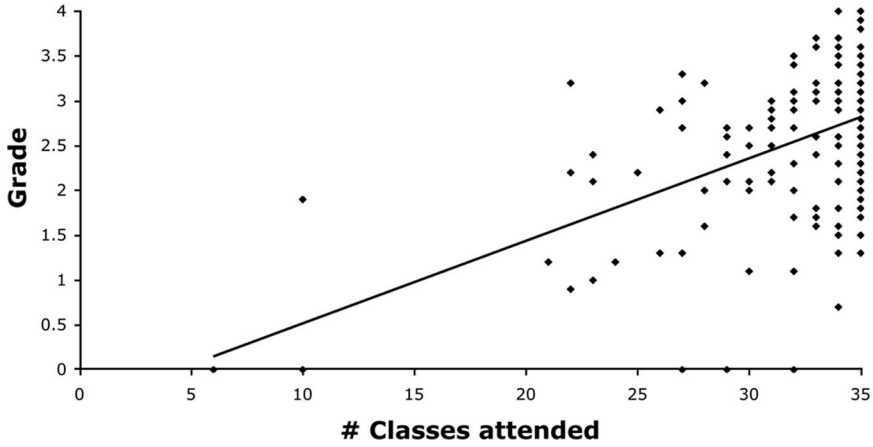

Figure 4. Class attendance is a significant predictor of final grade ($R^2 = 0.24$, n = 173, $p < 0.0001$).

3B). A regression analysis indicated that attendance was again a significant predictor of final grade ($R^2 = 0.18$, n = 310, $p < 0.0001$).

DISCUSSION

This study is one of many showing that active learning increases student performance in undergraduate science courses (e.g., Ebert-May *et al.*, 1997; Mazur, 1997; Crouch and Mazur, 2001; Knight and Wood, 2005; McConnell *et al.*, 2006). The unique aspects of this research are the emphasis on grading or public display of cards as a way to prescribe student participation in active-learning exercises and the ability to distinguish performance by high- and low-risk students. The fundamental messages of this work are that introductory biology students benefit from highly structured active-learning environments and that highly structured course designs may have a particular benefit for students who are at high risk of failing the course. At our university, high-risk students include a disproportionate number of URM and EOP individuals. Therefore, our findings provide some insight for improving introductory biology courses in a way that will help retain students who are historically underrepresented in science majors.

Risk Analysis

The risk analysis showed that UW GPA and SAT verbal are the most reliable predictors of student grades in the initial course in our introductory biology sequence. Thus, students who struggle in the initial biology course are also struggling in other college courses, and students with poor verbal skills do poorly on the course's written midterm and final exams.

Table 4. Predicted grades in Biology 180, fall 2005

Participation	2.77 ± 0.57 (154)
Graded	2.59 ± 0.61 (171)

Values are mean ± SD, with n in parentheses. *t* test, $p = 0.007$.

It is interesting to note that when we analyzed the overall risk of failing to complete the entire three-quarter sequence successfully, the most reliable predictors were UW GPA at the time of entering each course and SAT math score. We interpret the change from SAT verbal as an initial predictor to SAT math as an overall predictor as follows: Students with poor verbal skills quickly fail to progress in the series, whereas analytical and quantitative skills become more important over the combined sequence of courses. It would be interesting to know whether college GPA and SAT verbal or math scores are equally robust predictors of performance in introductory biology courses at other institutions.

Spring 2005 Course Design

The spring 2005 course designs were inspired by an observation we made during the spring 2003 course: initially, students participated enthusiastically in ungraded active-learning exercises. As the course wore on, however, participation dropped dramatically. In effect, many students appeared to say to themselves, "I'm not being graded on this stuff and I've got an organic chemistry midterm tomorrow; I'm not going to bother."

The four course designs tested in spring 2005 attempted to address the general issue of prescribing or "enforcing" active learning. More specifically, the course designs focused on addressing the four hypotheses proposed to explain the traditionally high failure rate. The goal of the weekly practice exams was to provide ESL and other students with

Table 5. Performance on common exam questions, fall 2005

	All students[a]	High-risk students only[b]	Low-risk students only[c]
Participation	229 ± 39 (154)	191 ± 35 (50)	247 ± 25 (104)
Graded	218 ± 43 (171)	187 ± 34 (71)	240 ± 33 (100)

Values are total points (mean ± SD), with n in parentheses.
[a] One-way ANCOVA, NS.
[b] *t* test, NS.
[c] *t* test, NS.

Table 6. Performance on clicker questions, fall 2005

	All students[a]	High-risk students only[b]	Low-risk students only[c]
Participation	55.2 ± 11.1 (147)	46.9 ± 11.0 (45)	58.9 ± 9.0 (102)
Graded	58.6 ± 11.1 (159)	51.8 ± 10.5 (71)	62.8 ± 9.3 (97)

Values percent correct (mean ± SD), with n in parentheses.
[a] One-way ANCOVA, $p < 0.001$.
[b] t test, $p = 0.021$.
[c] t test, $p = 0.003$.

increased opportunities to practice writing exam questions posed at the application and analysis levels of Bloom's taxonomy. The goals of the daily in-class questions were to provide additional practice with answering questions at relatively high levels of Bloom's taxonomy and to encourage active participation in class through grading or public declaration of answers. We predicted that the combination of daily and weekly graded practice would raise student effort level and help sustain it throughout the course.

The dramatic gains in student achievement in all of the spring 2005 course designs support the hypothesis that introductory students benefit more from active learning exercises when they are prescribed or enforced in some way as opposed to being voluntary and not associated with earning course points. If other research supports this conclusion, it will challenge instructors to design active-learning environments where students see immediate consequences if they fail to participate.

It is important to consider why study groups and individual practice worked equally well in raising student achievement in our course, even though study groups have been shown to raise student performance in other contexts in undergraduate science courses (Born *et al.,* 2002; Cortright *et al.,* 2003; Zeilik and Morris, 2004; Peters, 2005; Sharma *et al.,* 2005). We hypothesize that in this case, individual practice was just as effective as peer interaction because our practice exercises focused so narrowly on written exam questions. In this context, individual work may have had a "close-to-the-real-thing" benefit in providing exam practice that balanced out the benefits that peer interaction may have had in study groups.

This study adds to a growing literature on the efficacy of using electronic response devices in class. It is important to emphasize, however, that answering in-class questions with cards instead of clickers worked equally well in our course except for high-risk students, where our data indicate that students using clickers may marginally outperform students using cards. Both cards and clickers are used routinely in other courses on our campus and both have been shown to increase student achievement on exams in introductory science courses (e.g., Meltzer and Manivannan, 2002; Byrd *et al.,* 2003). Using clickers increased attendance primarily because it allowed points to be assigned, and increased attendance may have contributed to higher achievement by high-risk students. Although cards may work equally well for high-risk students if other mechanisms are in place to promote high attendance, the public nature of card use may lower the performance of high-risk students by contributing to stereotype threat (Steele, 1997). In this case, stereotype threat would occur if members of URMs felt that instructors and peers expected them to do poorly on public responses to in-class questions.

Fall 2005 Course Design

The fall 2005 course designs were inspired by the observation that high-risk students may benefit from course designs that include daily use of clickers. In addition, individual practice with written exam questions worked equally well as study group practice and was easier for us to administer. It is important to note that the increase in total exam points observed in spring 2005 was replicated in fall 2005, suggesting that prescribed active-learning course designs are robust in terms of raising student performance.

We compared right/wrong versus participation grading schemes for clicker questions to assess whether simply increasing attendance alone can be responsible for raising student achievement on exams. As predicted, both approaches were successful in promoting attendance to an average of 95%. Grading also clearly increased student achievement on the in-class questions themselves. However, because there was no difference in exam scores between the sections, we have no evidence of a carry-over from increased performance on in-class multiple-choice questions to increased performance on written exams. We are still unsure how much of the increase in exam performance is simply due to increased class attendance.

Low-risk students in the fall course did not perform better on common exam questions in the participation section, even though there were significantly more low-risk students in that section. This observation is important: The fall trial did not replicate the result from the spring course designs, where low-risk students did better if they were in a section with a significantly greater percentage of low-risk students.

In both spring and fall, there was a positive relationship between class attendance and overall course grade. Our data are consistent with other recent work showing that coming to class helps performance (Thomas and Higbee, 2000; Moore, 2003). We are unsure exactly what has to happen in class for this benefit to occur, however. Our data suggest that simply being in class and responding to questions has a benefit for students.

In evaluating data on clickers (and cards), it is also critical to note *how* the devices are being used. In all of our course designs, in-class questions were posed in a highly structured format. Each class session started with questions designed to reward students who had reviewed the previous day's material and done the reading for the day. The question on the reading was posed at the knowledge or content level, but all other questions were posed at the comprehension, application, or analytical level. Students were allowed to discuss questions with peers and reanswer if the initial percentage of correct answers was low, and the instructor frequently encouraged postquestion discussion on why answers were correct or incorrect. As the literature on the use of in-class questions matures, it will be important to assess the questioning structure and question types that are posed in order to compare and evaluate the costs and benefits of this aspect of course design.

To summarize, our data show that prescribed active learning benefits students in introductory biology more than voluntary or "unenforced" active learning. This study also shows that if introductory science courses are reformed in a way that prescribes constant student participation and practice, it is likely that more students, especially those who are at high risk of failing, will gain the discipline and intellectual tools required to be successful in the life sciences.

ACKNOWLEDGMENTS

We are deeply grateful to the students in Biology 180 for their advice on ways to improve the course and for their encouragement and support of this study. The risk analysis, course staffing, and purchase of a clicker system were funded by a grant from the University of Washington's College of Arts and Sciences as part of the Foundations Course program. We thank Biology Department Chair Tom Daniel for writing this grant. Data analysis was supported by a grant from the University of Washington–Howard Hughes Medical Institute Undergraduate Science Education Programs (Grant 52003841). We thank Michael Griego, Cathy Beyer, and Deb McGhee of the UW Office of Educational Assessment for developing the risk analysis and risk predictor model.

REFERENCES

Angelo, T. A., and Cross, K. P. (1993). Classroom Assessment Techniques, 2nd ed., San Francisco: Jossey-Bass.

Bloom, B. S. (ed.) (1956). Taxonomy of Educational Objectives: The Classification of Educational Goals, by a Committee of College and University Examiners, New York: Longmans Green.

Born, W., Revelle, W., and Pinto, L. (2002). Improving biology performance with workshop groups. J. Sci. Educ. Tech. *11*, 347–365.

Boyd, B. L. (2001). Formative classroom assessment: learner focused. Ag. Educ. Mag. *73*, 18–19.

Byrd, G. G., Coleman, S., and Wernath, C. (2003). Exploring the universe together: cooperative quizzes with and without a classroom performance system in Astronomy 101. Astron. Educ. Rev. *3*, 26–30.

Cech, T., and Kennedy, D. (2005). Doing more for Kate. Science *310*, 1741.

Cortright, R. N., Collins, H. L., Rodenbaugh, D. W., and DiCarlo, S. E. (2003). Student retention of course content is improved by collaborative-group testing. Adv. Physiol. Educ. *27*, 102–108.

Cota-Robles, E. H., and Gordan, E. W. (1999). Reaching the Top: A Report of the National Task Force on Minority High Achievement, New York: College Board.

Crouch, C. H., and Mazur, E. (2001). Peer instruction: ten years of experience and results. Am. J. Phys. *69*, 970–977.

Dirks, C., and Cunningham, M. (2006). Enhancing diversity in science: is teaching scientific process skills the answer? CBE Life Sci. Educ. *5*, 218–226.

Ebert-May, D., Brewer, C. A., and Allred, S. (1997). Innovation in large lectures—teaching for active learning. BioScience *47*, 601–607.

Gandara, P., and Maxwell-Jolly, J. (1999). Priming the Pump: Strategies for Increasing Underrepresented Minority Graduates, New York: College Board.

Handelsman, J. *et al.* (2004). Scientific teaching. Science *304*, 521–522.

Knight, J. K., and Wood, W. B. (2005). Teaching more by lecturing less. Cell Biol. Educ. *4*, 298–310.

Lyman, F. T. (1981). The responsive classroom discussion: the inclusion of all students. In: Mainstreaming Digest, ed. A. Anderson, College Park: University of Maryland Press, pp. 109–113.

Matsui, J., Liu, R., and Kane, C. M. (2003). Evaluating a science diversity program at UC Berkeley: more questions than answers. Cell Biol. Educ. *2*, 117–121.

Mazur, E. (1997). Peer Instruction: A User's Manual, Upper Saddle River, NJ: Prentice Hall.

McConnell, D. A. *et al.* (2006). Using conceptests to assess and improve student conceptual understanding in introductory geoscience courses. J. Geosci. Educ. *54*, 61–68.

Meltzer, D. E., and Manivannan, K. (2002). Transforming the lecture-hall environment: the fully interactive physics lecture. Am. J. Phys. *70*, 639–654.

Moore, R. (2003). Attendance and performance: how important is it for students to attend class? J. Coll. Sci. Teach. *32*, 367–371.

Mosteller, F. (1989). The "Muddiest Point in the Lecture" as a feedback device. On Teach. Learn. J. Harvard-Danforth Cent. *3*, 10–21.

Peters, A. (2005). Teaching biochemistry at a minority-serving institution: an evaluation of the role of collaborative learning as a tool for science mastery. J. Chem. Educ. *82*, 571–574.

Sharma, M. D., Mendez, A., and O'Byrne, J. W. (2005). The relationship between attendance in student-centred physics tutorials and performance in university examinations. Int. J. Sci. Educ. *27*, 1375–1389.

Steele, C. M. (1997). A threat in the air: how stereotypes shape intellectual identity and performance. Am. Psychologist *52*, 613–629.

Summers, M. F., and Hrabowski, F. A., III. (2006). Preparing minority scientists and engineers. Science *311*, 1870–1871.

Thomas, P. V., and Higbee, J. L. (2000). The relationship between involvement and success in developmental algebra. J. Coll. Read. Learn. *30*, 222–232.

Zeilik, M., and Morris, V. J. (2004). The impact of cooperative quizzes in a large introductory astronomy course for non-science majors. Astron. Educ. Rev. *3*, 51–61.