CBE—Life Sciences Education
Vol. 5, 270–280, Fall 2006

Article

Rethinking Exams and Letter Grades: How Much Can Teachers Delegate to Students?

Elizabeth Kitchen,* Summer H. King,† Diane F. Robison,‡ Richard R. Sudweeks,‡ William S. Bradshaw,* and John D. Bell†

Departments of *Microbiology and Molecular Biology, †Physiology and Developmental Biology, and ‡Instructional Psychology and Technology, Brigham Young University, Provo, UT 84602

Submitted November 9, 2005; Accepted March 22, 2006
Monitoring Editor: Gary Reiness

In this article we report a 3-yr study of a large-enrollment Cell Biology course focused on developing student skill in scientific reasoning and data interpretation. Specifically, the study tested the hypothesis that converting the role of exams from summative grading devices to formative tools would increase student success in acquiring those skills. Traditional midterm examinations were replaced by weekly assessments administered under test-like conditions and followed immediately by extensive self, peer, and instructor feedback. Course grades were criterion based and derived using data from the final exam. To alleviate anxiety associated with a single grading instrument, students were given the option of informing the grading process with evidence from weekly assessments. A comparative analysis was conducted to determine the impact of these design changes on both performance and measures of student affect. Results at the end of each year were used to inform modifications to the course in subsequent years. Significant improvements in student performance and attitudes were observed as refinements were implemented. The findings from this study emphasized the importance of prolonging student opportunity and motivation to improve by delaying grade decisions, providing frequent and immediate performance feedback, and designing that feedback to be maximally formative and minimally punitive.

INTRODUCTION

A recent trend in biology education has been to move from the traditional dissemination of a large volume of encyclopedic facts toward an emphasis on the acquisition of scientific thinking skills. In 1996, the National Research Council (NRC) published the *National Science Education Standards*, which portrayed a vision of a scientifically literate society. It proposed a shift from a traditional science education to one based on active learning in which students participate in problem solving, apply knowledge to new situations, ask questions, and make informed decisions (Bonwell and Eison, 1991; NRC, 1996). Although the ideas presented in the *National Science Education Standards* were not new, the report by the NRC opened a national dialogue and encouraged reform throughout the educational system. Nevertheless, progress toward successful implementation has been slow. Traditional pedagogical practices may not be perfectly suited to this new emphasis on acquiring analytical thinking skills (Rifkin and Beorgakakos, 1996; Beyer, 2001; Hay, 2001).

The "assessment environment" (term coined by Stiggins and Conklin, 1992) is one component of education that we believe deserves scrutiny, particularly when creating courses designed to promote problem solving. Acquisition of scientific or analytical thinking skills can be a challenging process for students, one often fraught with periods of failure and discouragement. For example, in our experience, the "grade threat" can loom large under the traditional grading system as students struggle toward mastery, causing them to be reluctant to take the risks necessary to acquire such skills. The acquisition of scientific or analytical thinking skills involves working in authentic situations with problems that are often nebulous or underdefined (Huba and Freed, 1999, Chapter 2). Students will make errors when

DOI: 10.1187/cbe.05–11–0123
Address correspondence to: John D. Bell (John_Bell@byu.edu).

attempting to solve such problems. In fact these errors are often a helpful component of the learning process. "In learner-centered teaching. . . mistakes are opportunities on which to capitalize, rather than events to avoid" (Huba and Freed, 1999, p. 46). If failure is punished with poor grades, students may become discouraged. We have noticed that this frustration is commonly expressed in terms such as "I studied even harder for this exam; I still did poorly; I can't get any better." From our perspective, we believe that what the student really means is "I don't know how to improve." In general, it appears that students who express these frustrations choose one of three ineffective methods of coping: 1) they exert more effort repeating unsuccessful strategies used previously; 2) they blame external factors (the test, the teacher, the text); or 3) they simply quit. This study addressed these issues by examining ways to motivate students and track their progress by providing feedback that is 1) frequent, 2) nonthreatening, and 3) formative, while alleviating frustrations associated with grading.

Frequent Feedback

A hallmark of student-centered course design is the use of frequent assessment to provide both student and instructor with a measure of student achievement and comprehension of course concepts (Huba and Freed, 1999; NRC, 2000; Weimer, 2002). Models of active-learning pedagogy include frequent assessments in the form of daily quizzes or short exams that cover small units of instruction in addition to the use of in-class discussion or group problem-solving activities (Klionsky, 2001; McConnell *et al.*, 2003; Burrowes, 2003; O'Sullivan and Copper, 2003; Fitzpatrick, 2004). L. Dee Fink uses the acronym FIDeLity to summarize his view of high-quality feedback: Frequent, Immediate, Discriminating, and done Lovingly (Fink, 2003). The obvious logic of this approach is that providing students multiple opportunities to complete a learning cycle (attempt, receive feedback, make decisions for improvement, attempt again) will enhance the probability of developing analytical thinking skills. Because it is the iterative nature of the process that seems valuable, changes in course design that increase the number of formative episodes will likely be profitable. This logic is supported by John A. Glover's study supporting the "testing phenomenon" (Glover, 1989). Students who are tested on material between the first time they study it and the final test remember more of the material than students who do not take an intervening test.

Nonthreatening Feedback

Traditionally, in science, most assessments have taken the form of graded tests. This practice has come under recent scrutiny of late because of increasing interest in the connection between students' affective characteristics and learning. Kohn (2004) identified three negative outcomes associated with grading: 1) reduction of student interest in learning, 2) reduction in student preference for challenging tasks, and 3) reduction in quality of student thinking.

We would add to this list that issuing grades without helping students understand the criteria used does little to develop students' ability to self-assess and diagnose. Development of higher-level cognitive skills, such as those assessed in scientific problem solving, requires explicit diagnostic feedback (Huba and Freed, 1999).

A number of studies have supported Kohn's views, showing that grading as opposed to nonthreatening, formative feedback tends to result in lowered performance levels, decreased positive affect, or both (for example, Benware and Deci, 1984; Hughes *et al.*, 1985; Butler and Nissan, 1986; Black and Wiliam, 1998). Harmful aspects of grading may be magnified in classroom situations where difficult, higher-level cognitive skills are taught. In a study by Hughes *et al.* (1985) a significant interaction was found between task difficulty level and evaluative condition. If the cognitive task was simple, the negative impact of teacher-imposed evaluation on student affect was modest. In contrast, when the task was difficult, the impact was much greater.

Formative Feedback

In order for students to self-assess and manage their own learning, they need frequent snapshots of their status in relation to mastery standards. "Thus self-assessment by pupils, far from being a luxury, is in fact *an essential component of formative assessment*. When anyone is trying to learn, feedback about the effort has three elements: recognition of the *desired goal*, evidence about *present position*, and some understanding of a *way to close the gap* between the two" (Black and Wiliam, 1998, p. 10; authors' emphasis). We have observed that students commonly lack the ability to monitor and evaluate their own progress in a course and to make informed decisions for improving their learning. We believed that developing this skill must therefore be a central part of our pedagogical experiment. We based this work on the premise that the evaluation process chosen for a course may exert a substantial influence on the types of learning strategies that students adopt and the success they achieve (Crooks, 1988; NRC, 2001). In an extensive review article, Black and Wiliam drew the conclusion that:

> There is a firm body of evidence that formative assessment is an essential component of classroom work and that its development can raise standards of achievement. We know of no other way of raising standards for which such a strong *prima facie* case can be made.
>
> (Black and Wiliam, 1998)

Accordingly, our objectives were to 1) replace traditional midterm exams with short weekly formative assessments; 2) create formal means for students to obtain immediate feedback on their performance from themselves, their peers, and their instructor; and 3) involve the students in the process of assigning grades. During a three-semester study incorporating formative feedback into our Cellular Biology course, the most clear and important lesson we learned was that the best gains in both affect and performance resulted from strategies that maximized feedback and minimized the impact of grades. We also learned that students are accepting of a novel approach such as this; they generally perceive that this format provides better feedback, but demonstrating objectively that this format leads to improved performance is more complex than anticipated and requires iteration and ongoing evaluation of the implementation strategies.

MATERIALS AND METHODS

Course Design

Biology 360 is a large-enrollment, three-credit, upper-division Cellular Biology course that meets three times a week for 50 min per session. It is a required course for students majoring in basic life sciences at Brigham Young University. Prerequisites for the course include a two-credit Molecular Biology course, a two-credit Genetics course, and Organic Chemistry. The texts used in the course during this study were two editions of *Molecular Cell Biology* (Lodish *et al.*, 2000, 2003). The instructor for the 3 yr of testing the formative version of the course was John Bell, one of the authors of this article. Course content is divided into 11 topics (Table 1). The primary focus of the course is to apply this content to solving scientific problems. These problems consist of a short narrative about a modern experiment, and students are expected to draw defensible conclusions from the accompanying tabular or graphical data (Kitchen *et al.*, 2003).

The formative version of the course described in this article involved a scheme in which the content topics were aligned with a weekly layout of three distinct class sessions. The first session of the week (generally Mondays) focused on the conceptual components of that week's topic. The second session (Wednesday) was a problem-solving workshop that offered directed practice in applying the concepts learned to the interpretation of experimental data. The final session each week (Friday) was reserved for formative assessment. In the three cases in which the topic spanned a 2-wk period, the three-session cycle was repeated during the second week.

Mondays were devoted to mastering conceptual models. Students were expected to come to class with an understanding of the topic so that class time could be spent processing the information instead of reviewing the text in a traditional lecture. Accordingly, they were assigned to read ~40 pages of text before class. The classroom agenda consisted of learning activities designed to maximize student involvement by focusing on clarification from the instructor, diagramming cellular processes, teaching concepts to peers, and formulating questions.

The Wednesday problem-solving workshop offered students directed practice in solving data analysis problems relevant to the week's topic. Students worked collaboratively to interpret data and write justifiable conclusions. The instructor circulated among the students, monitoring the process and offering assistance. Students gained further practice with homework problems of similar design. The instructor encouraged them to work these assignments in small groups. Students were provided sample answers to the problems to allow them to assess their understanding. Completion of the assignments was self-reported. Students were also given the opportunity to meet with teaching assistants or the instructor during office hours.

Table 1. Course topics for Biology 360

Week	Topic
1	Energy, equilibrium
2	Regulation of protein function
3	Immunology
4	Membrane biochemistry
5	Organelles and the secretory pathway
6	Mitochondria and chloroplasts
7	Cytoskeleton
8 and 9[a]	Transcriptional regulation in eukaryotes
10	Cell cycle control
11 and 12[a]	Signal transduction
13 and 14[a]	Embryonic development

[a] These topics spanned 2 wk rather than 1 wk.

The weekly performance assessments on Fridays consisted of one to three conceptual problems and one data analysis problem (see Supplemental Material). Conceptual problems ranged in difficulty. The simplest problems asked students to provide a definition. The more challenging items required students to diagram major components of the current topic. Data analysis problems instructed students to read a passage and write valid conclusions based on graphical or tabular data as described above. To simulate the test-like conditions of an exam, students worked silently and individually. After 25 min, the instructor presented samples of optimal answers and facilitated student discussion to help students recognize appropriate conclusions. Students worked in pairs to rate their own assessments for the remainder of the 50-min period. In addition, they identified and recorded ways to improve future responses. The instructor participated actively with students in this process. During the third year of the study an optional hour for additional assistance was available immediately after the assessment session.

As explained in *Results*, the details of this scheme evolved in response to student performance and results of student surveys over the three semesters. The details of these changes and the reasoning behind them are summarized in Table 2.

Scoring of Performance Assessments and Grading

The final exam was used as the summative assessment for the course both for assignment of letter grades to students and for evaluation of this study. It consisted of conceptual and data analysis items. The data analysis problems from the final exam, although similar in format to problems from the weekly assessments, utilized data sets and experiments that the students had not encountered previously. Thus, improvement in student performance between the weekly assessments and the final exam should be attributed to gain in analytical reasoning ability rather than familiarity with the problems (sample items are in the Supplemental Material). For the purposes of this study, data analysis problems on final exams were scored twice using different raters. Where there was disagreement between the original two ratings on a given item, a third rater also scored that item and the three ratings were averaged.

To promote cooperation rather than competition among students and to provide clarity for self-assessment, the grading scheme for the course was criterion based. The following criteria were used for assigning grades:

- "D" level: ability to answer simple recall-type questions
- "C" level: mastery of "D" level plus mastery of course concepts as indicated by the ability to draw or explain cellular processes
- "B" level: mastery of "C" level plus ability to apply concepts to novel contexts
- "A" level: mastery of "B" level plus mastery of data analysis problems

In the second and third years, the "D"- and "B"-level problems were abandoned, and those grades were given to students who displayed intermediate levels of performance. Thus, students who were able to answer "C"-level problems only partially or half of the time but never succeed with "A"-level problems received a "D" grade. Those who could consistently answer "C"-level problems but could answer "A"-level problems only partially or half of the time received a "B."

During the course, students were taught to apply this rubric from the final exam to rate their weekly assessments. This scheme was critical in order for students to evaluate their progress in the course. Assessments were collected by the instructor after scoring and kept on file for future reference (see next paragraph). To estimate the accuracy of student self-rating, a teaching assistant rescored the weekly assessments. The teaching assistant was trained by the instructor. Accuracy of scoring was validated by comparing the teaching assistant's ratings of student responses to ratings generated by course instructors. During the second year of the study, students were apprised of the score assigned by the teaching assistant (see

Table 2. Course alterations during the 3 yr of the study

	Component of class	Original implementation
Original course	Reading assignment	Information acquisition prior to class Accountability achieved with quiz
	Homework	Assigned weekly Data analysis items
	Midterm exams	Administered every 3 wk Data analysis and conceptual items
	Grading system	Exam items scored by teaching assistants Final score aggregated from midterm and final exams
	Objective	**Change**
Year 1	Create a course structure that replaces midterm exams with weekly assessments	Consolidated set of weekly topics into one unit Dedicated Mondays to clarification of concepts Dedicated Wednesdays to practicing data analysis Dedicated Fridays to performance assessment
	Adjust reading assignments to new weekly structure	Consolidated according to weekly topic Completion required prior to Monday each week Accountability based on self-reporting
	Create a formative culture	Replaced midterm exams with weekly assessments Feedback provided immediately following assessment Student pairs evaluated their assessments relative to optimal examples
	Adjust grading system to reflect desired culture	Grading/scoring system criterion based Final exam scored by instructor and used as initial basis for assigning grades Students proposed a course grade defended by data from weekly assessments Instructor adjusted final grades when justified
Year 2	Increase student compliance with reading assignments	Conceptual portion of weekly assessments administered on Monday prior to class discussion Provided 5-min reading overview starting week 6
	Improve accuracy of student self-scoring	Simplified rubric Assessments re-scored by teaching assistants and the corrected scores reported to students
Year 3	Focus attention on meaningful feedback for improvement	Conceptual and data analysis items administered together on Fridays Added voluntary hour for performance feedback after Friday's class Abandoned correction of student self-scores
	Maintain benefits of year 2	Retained simplified grading rubric Retained 5-min reading overview

Table 2). This feedback was intended to help students calibrate their level of performance in the course. However, as explained in *Results*, this practice was abandoned during the third year.

To mitigate anxiety concerning use of the final exam as the sole summative evaluation, students were informed at the beginning of the semester that they would have the option at the end of the course of choosing to include weekly performance data for consideration in their course grades. At the time of the final exam, students exercising this option wrote a one-page proposal identifying their perception of an appropriate course grade. Students were expected to use the data from their weekly assessments to determine and justify their proposed grade. When the proposed grade was higher than that calculated from the final exam, the instructor evaluated the justification and data from the student's file of collected assessments. All of this information was then used to negotiate a fair course grade. Approximately 90% of the students submitted proposals.

Affective Assessments

Student attitudes were assessed using a questionnaire administered as a take-home assignment during the last 2 wk before the final exam. The questionnaire contained 41–44 questions, depending on the year. Forty questions remained constant for all three years, and the others varied in order to collect data on course changes specific to a given year. Students were asked about their perceptions of the quality of the course, its impact on continuing interest in science and cell biology, the usefulness of the grading techniques, the quality of feedback received, self-efficacy, and improvements in their abilities. (Complete survey and results are available on request from the corresponding author.)

In addition, students had the option of completing an online survey administered by the university. It was offered during the last 3 wk of the semester before final exams were administered. For this study, data regarding student out-of-class preparation time and Likert-type ratings of instructor, course, exam, and grading quality were extracted from the survey results.

Statistical Analysis

Descriptive data were gathered to compare the populations of students taking the course during the three semesters. The data col-

Table 3. Comparison of class demographics during the 3 yr of the study

Comparison criterion	Original	Formative			p value[a]
		Year 1	Year 2	Year 3	
Female representation in class (%)	30	16	22	16	
Grade in prerequisite course (4.00 scale)	2.82 ± 0.10[b]	2.84 ± 0.10	3.10 ± 0.10	2.97 ± 0.09	0.20
Credits previously earned	118 ± 3.0	126.0 ± 2.6	121.7 ± 2.6	120.4 ± 2.7	0.30

[a] One-way ANOVA.

[b] Mean ± SEM.

lected were the number of credit hours students earned before enrollment in Biology 360, their grade in a prerequisite Molecular Biology course, and their gender. Comparisons of mean credit hours and grades were achieved by analysis of variance (ANOVA). As shown in Table 3, there were no significant differences in these parameters among the 3 yr or in a parallel section ("Original") taught in the former design during year 1. Some variation in gender distribution was observed among the various groups (Table 3). Nevertheless, no difference in performance was detected between male and female students in any of the groups (by t test, $p > 0.2$ in each case). Whether gender contributed to the results of affective surveys cannot be ascertained because the surveys were completely anonymous.

Differences in the percentages of class members choosing the various responses to items on the anonymous questionnaire (Table 4) were analyzed using a chi-square test of independence. In this version of the chi-square test, the null hypothesis states that the percentages for any row of the table are the same; hence, the expected values are the computed means for the data across each row. For our data, the test determines whether there is a statistically significant difference in the distribution of responses to each item between two or more of the semesters examined (including both original and formative trials).

Trends during Friday assessments within a semester were analyzed by linear regression. Comparisons of performance and other continuous data between semesters were analyzed by two-tailed t tests or ANOVA as appropriate. Answers to the affective assessment were analyzed by chi-square. Data are expressed as the mean ± SD or SEM as indicated and appropriate.

RESULTS

Year 1

Rationale. For several years, we had experimented with methods to teach skills in scientific reasoning in our cell biology course (Kitchen *et al.,* 2003). In recent years, we established an effective design that produced consistent levels of student performance from semester to semester regardless of instructor (J.D.B. or W.S.B.; $p = 0.38$ by ANOVA, n = 4 semesters from 1999 through 2001). In an attempt to further elevate performance, we initiated trials of a new "formative" assessment format in which midterm exams were replaced with weekly short assessments and a modified grading scheme as outlined in *Materials and Methods.* Other than this modification, the original and formative versions of this course were very similar. The cognitive objectives, subject matter topics, depth of coverage, in-class practice exercises, and analytical problems used as assessment instruments were essentially identical. One unique feature of the original format was that, in addition to the three regular class periods per week, each student attended an additional small-group mentoring session that was devoted to solving homework problems under the tutelage of the instructor (W.S.B.). The reason that these mentoring sessions were not included in the formative version of the course was that we believed that the Wednesday workshops (see *Materials and Methods*) provided an equivalent experience. As described below for year 3, however, we did eventually discover that additional scheduled time with the instructor was beneficial.

Outcomes. Student attitudes toward the formative format during year 1 appeared to be positive (Table 4). In most cases, responses to survey questions were comparable to those obtained in a parallel section of the course taught in the original format, although some differences were observed. Notably, students felt that grading procedures were fair and that weekly assessments were useful for learning. Students reported that they improved their data analysis skills during the course and that the course format was desirable. Interestingly, students also reported that they were able to focus more on learning and worry less about their grades than in a more traditional classroom setting.

Performance on data analysis tasks was indistinguishable from that observed in a parallel section of the course taught in the original format ($p = 0.49$ by t test, Figure 1). We had hoped that the formative version would lead to significant gains in performance. Because this did not occur, we sought information from the data that might inform potential improvements to this format and perhaps produce enhanced student performance. One clue came from the observation that students in the formative section spent 1.7 fewer hours outside of class preparing for the course (original version: 6.5 ± 0.3 [SEM] hours, formative version: 4.8 ± 0.3 h, n = 52–64, $p < 0.0001$ by two-tailed t test). We suspected that this tendency could reflect a false sense of security and diminished motivation due to a lack of the accountability and sobering appraisal inherent to traditional midterm exams. This interpretation was substantiated by evidence that students overestimated their performance when asked to report and defend their understanding of the grade level they were achieving in the course (Figure 2A).

Year 2

Rationale. We acted on our findings from the first year by effecting three concrete changes to the format for year 2

Table 4. Sample of responses to student survey

Item	Question and responses	Original (n = 64)	Formative Year 1 (n = 62)	Formative Year 2 (n = 49)	Formative Year 3 (n = 91)	*p* value[a]
7	How do you rate the fairness of the grading procedures and criteria used in this course?					0.00005
	Completely fair	10.9	22.6	12.5	31.5	
	Quite fair	50.0	61.3	35.4	50.6	
	Moderately fair	29.7	16.1	33.3	16.9	
	Not very fair	7.8	0	16.7	1.1	
	Not fair at all	1.6	0	2.1	0	
8	In your judgment, to what degree was the curriculum of this course appropriately distributed in terms of the amount of emphasis placed on learning the subject-matter content vs. the emphasis placed on learning the thinking skills and analytical methods taught in the class? The course placed					0.0003
	Too much emphasis on the subject-matter content	0	4.8	0	0	
	About the right amount of emphasis on both subject matter and thinking/analytical skills	75.0	87.1	75.5	93.4	
	Too much emphasis on the thinking skills and analytical methods	25.0	8.1	24.5	6.6	
14	Have you in fact made improvement in the ability to interpret experimental data as a result of your participation in this course?					0.00004
	Definitely yes	48.4	51.6	38.8	76.9	
	Probably yes	31.3	45.2	44.9	20.9	
	Probably not	9.4	1.6	10.2	2.2	
	Definitely not	3.1	0	0	0	
	I'm not sure	7.8	1.6	6.1	0	
35	In this system I have been able to focus my attention more on learning and worry less about my grade.					0.00009
	Strongly agree	NA	56.5	28.6	72.1	
	Slightly agree	NA	27.4	26.5	20.6	
	Neither agree or disagree	NA	8.1	20.4	1.5	
	Slightly disagree	NA	4.8	16.3	2.9	
	Strongly disagree	NA	3.2	8.2	2.9	
38	The weekly assessments have helped me to improve my data analysis skills more effectively than would have been possible in a more traditional exam setting					0.0002
	Strongly agree	NA	72.6	34.7	76.9	
	Slightly agree	NA	21.0	40.8	12.1	
	Neither agree or disagree	NA	3.2	12.2	6.6	
	Slightly disagree	NA	3.2	8.2	3.3	
	Strongly disagree	NA	0	4.1	1.1	
40	How do you assess the quality of feedback (knowing what you did well and what needed improvement) you have received during each Friday session?					0.000008
	I received higher quality feedback than I usually get with a traditional TA/Professor-graded exam	NA	72.6	42.9	83.5	
	I received poorer feedback than I usually get with a traditional TA/Professor-graded exam	NA	4.8	22.4	4.4	
	The quality of feedback I have received has been comparable to what I have received with a traditional TA/Professor-graded exam	NA	22.6	34.7	12.1	

Values are percentages. NA, not applicable.

[a] Statistical significance of response distributions determined by chi-square test of independence.

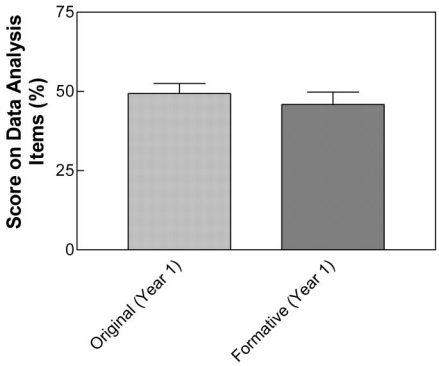

Figure 1. Comparison of student performance on data analysis tasks from the final exam in the original and formative formats of the course during year 1. Data represent average student scores on four data analysis items that were common between the final exams of the sections of the course taught in the original or formative formats. Items are scaled to a percentage of the points possible on those items. Data between the two groups are indistinguishable statistically based on a *t* test ($p = 0.49$, n = 70 "Original" or 69 "Formative").

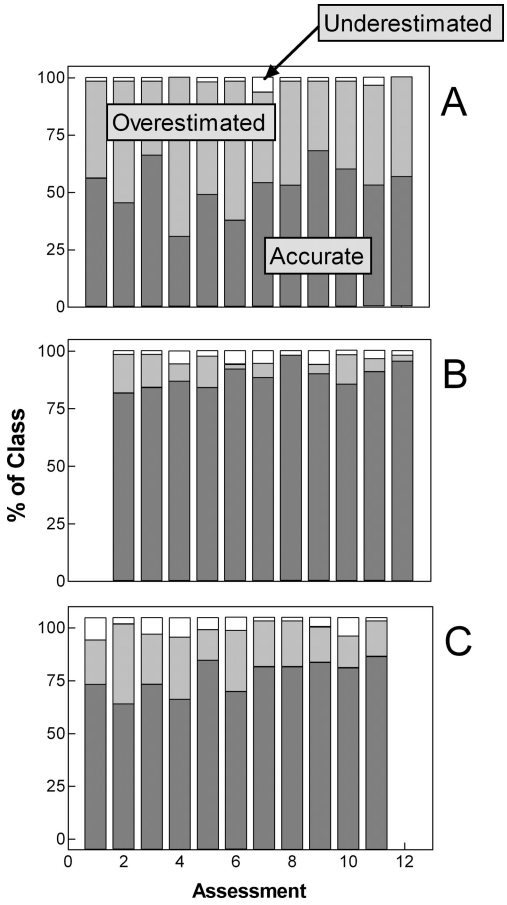

Figure 2. Self-grading accuracy on data analysis assessment problems in year 1 (A), year 2 (B), and year 3 (C). Student self-scores were compared with those assigned subsequently by the TA and classified as "accurate" (dark gray, student and TA agreed), "overestimated" (light gray, student score above TA's), or "underestimated" (white, student score below TA's). Trends in student accuracy of scoring across the weekly assessments were insignificant in A ($p = 0.37$) and significant in panels B and C ($p = 0.03$ and 0.009). Enrollments were 69 (year 1), 64 (year 2), and 97 (year 3).

(summarized in Table 2). First, to encourage more time in preparation, we separated the weekly assessments into two parts. The first part consisted only of conceptual problems (see *Materials and Methods*) and was offered during the first 15 min of class on Mondays. The second part, offered at the onset of class on Fridays, contained the data analysis problems. The purpose of this change was to make students accountable for reading and assimilating the basic information for that week's topic before Monday. In addition, we anticipated that the exercise would reveal to students their weaknesses and misconceptions, which would promote improved discussion and participation in class that day.

The second change was to simplify the grading rubric. We assumed that the disparity in student and instructor estimates of performance shown in Figure 2A was a consequence of complications in the grading scheme used during the first year. We reasoned that the letter grade communicated a clear message to students regarding their performance and that a more realistic view of that grade would provide stronger motivation to diligence and improvement.

The third alteration was also intended to help students better gauge their abilities. The teaching assistant (TA) provided revised assessment scores to students (see *Materials and Methods*) to validate and correct their self-scoring and help them more accurately monitor their progress. The summative assessment for official course grades, however, still depended solely on the final exam with student input as in year 1.

Outcomes. Two positive results indicated success of the format used during year 2. Most importantly, performance on data analysis tasks during the final exam improved by ~24% (Figure 3A). This improvement in year 2 seemed noteworthy because no variation in average performance had been observed previously across multiple offerings of the same course (see previous section). Moreover, the gain in student scores comparing performance on early assessments with that on the final exam was twice that observed during year 1 (Figure 3B). Another positive result is shown in Figure 2B. Student ability to self-evaluate and predict the appropriate letter grade during weekly assessments improved compared with year 1.

In contrast, Table 4 demonstrates that student attitudes were less positive; the high level of enthusiasm expressed during year 1 was not present during year 2. This change was particularly notable in questions referring to the assessment system.

One reason for these less-positive attitudes appeared to be frustration with the weekly assessments (items 37 and 38 in Table 4). Interestingly, students reported a lesser sense that feedback from assessments, a cornerstone in the formative format, was effective in helping them (item 40 in Table 4). As we explored the matter further, we learned that our effort to validate and correct self-scoring had focused their attention

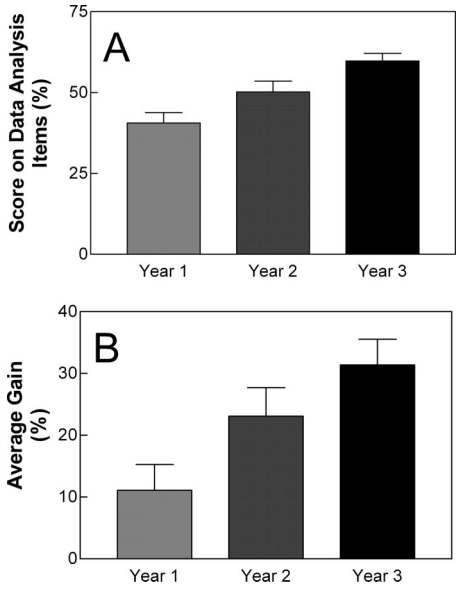

Figure 3. Student performance on data analysis tasks from the final exam in the formative format of the course. (A) Average performance on the six data analysis items of the final exam with scores scaled to a percentage of the points possible for those items. (B) Average percent gain in scores between data analysis items on the first two assessments in the course and those on the final exam. Error bars, SEM. Results for the various years were distinguishable statistically by ANOVA for both sets of data (A: $p < 0.0001$; B: $p = 0.004$, n = 64–97; see legend to Figure 2). Differences among groups were further analyzed by Dunnett's post test, and year 3 was distinguishable from year 1 in both cases at the level of $p < 0.01$.

too strongly on the letter grade per se (item 35 in Table 4). Consequently, it appeared that students' zeal to correctly identify their grade caused them to lose sight of the more valuable feedback designed to help them improve.

Ironically, the reported time spent outside class did not change between years 1 and 2. In year 2, the time spent outside class was 4.6 ± 0.4 (SEM) hours ($p = 0.75$ by t test). This became apparent to the instructor midway through the semester as he noticed that students seemed no better prepared for Monday's class in year 2 than in year 1. The instructor expressed frustration to the class for lack of diligence. One student reported privately that the problem was not laziness but rather difficulty in extracting the conceptual "big picture" from the information-rich text. After confirming the generality of this sentiment with the rest of the class, the instructor initiated midsemester a new practice of providing a brief (5-min) overview of the next reading assignment every Friday. This practice began week 6 and appeared to correspond to a modest surge in student performance on conceptual problems during weeks 6 through 12. The average conceptual score for weeks 1 through 5 was 64.4 ± 4% (SEM), whereas weeks 6 through 12 averaged 79.1 ± 5% ($p = 0.06$ by t test). In addition, students were asked at the end of the semester in the anonymous survey whether this innovation was helpful. All students reported that the feedback was moderately (16%) or very (84%) helpful.

Year 3

Rationale. The central objective for year 3 was to retain the improvement in student performance obtained during year 2 while restoring the positive nature of student affect observed in year 1. The plan to achieve this objective focused on maintaining realistic self-appraisal and engagement with the text before Mondays while redirecting attention away from the letter grade and back to useful feedback from the Friday assessments. Accordingly, we kept the 5-min reading overview and the simplified grading rubric. However, the feedback from the TA on grading accuracy was eliminated. The conceptual assessments given on Mondays during year 2 were returned to Fridays and administered together with the data analysis items. Lastly, an hour for additional feedback and discussion of the Friday assessments was added. Attendance during this extra session was voluntary.

Outcomes. Figure 3 demonstrates that student performance on data analysis tasks improved further by an additional 19% on the final exam (year 3 compared with year 2 in panel A), and the gain in scores was 36% higher than in year 2 (panel B). Thus, the overall increment from year 1 (in which performance was identical to the original version) to year 3 was 47% for end point performance (panel A) and 183% for score gain (panel B, significant by ANOVA followed by Dunnett's test, $p < 0.01$ in both cases). The average time spent outside of class was again 4.6 ± 0.2 (SEM) hours (n = 68; all three years indistinguishable by ANOVA, $p = 0.92$). Although students no longer received feedback regarding the accuracy of their self-scored weekly grades, Figure 1C reveals that their ability to accurately self-score was similar in year 3 compared with year 2.

Table 4 shows that student affect also improved in year 3 to be equal to or greater than that observed in year 1. The affective survey showed that students felt that the course was worth the effort they put in. In addition, students felt that they were able to focus on learning rather than their grade. Students felt grading procedures were fair and that they received a high quality of feedback. Attitudes about collaboration with others to rate assignments were favorable. Thus, it appeared that the adjustments made between year 2 and year 3 were beneficial.

DISCUSSION

Introduction of the formative course format resulted in both improved student performance and increased positive affect. In the original version of the course, student performance was already at a high level acceptable to instructors (Kitchen *et al.,* 2003). Moreover, this level had remained constant over several semesters, suggesting that the pedagogy used was consistent and well established. Thus, it was both exciting and significant to discover that further and large gains in performance could be realized by changing the system of examination and grading in the course. This improvement was complemented by the presence of positive student attitudes, especially during the third year. These outcomes developed over the 3 yr of this study as various aspects of the course were tailored to meet student needs as informed by evaluation data. During this process, a number

of key components were identified that had noticeable impact on favorable student performance and affect.

Communication and trust between the instructor and students were central facets of the course. Although the instructor had to trust that students would take the weekly assessments seriously, students had to trust that there would be alignment between the weekly assessments and the final exam. Frequent comments by students suggested that at first, they were uncertain they could achieve proficiency in data interpretation. The instructor had to provide encouragement and reasons for students to remain positive until they could see the benefits of their efforts. For instance, he informed them that students in past semesters had succeeded at data interpretation and that the process would become easier over time. The instructor reminded students that helping each other would not be penalized; rating was criterion based, and students were not competing with peers for a grade. Instructors attempting a similar course layout must recognize that trust can be a troubling issue for students at first.

Another important component of the course was control. We suspect from conversations with students that one source of frustration for students is the feeling they do not have control over a situation. Because students in this course were acquiring a new skill, they did not yet have command of data interpretation. We mollified student apprehension about the new technique by delaying judgment of performance in the course until the final exam, allowing students to achieve proficiency at their own pace. That is, a student who achieved "A"-level proficiency the day before the final exam would receive the same final grade as a student who worked at "A"-level from the first day of class. This format allowed time for students to experiment with different methods until they gained control over data interpretation and discovered what would lead to success. Weekly formative assessments also saved students the worry and anxiety often associated with midterm exams.

Some students may have been nervous at first that their grade was determined only by performance on the final exam. However, we found that students generally performed the same or better on the final exam than on the weekly assessments (for example, in year 3, 18% scored the same, and 57% scored higher on the final exam). In addition, students scored their own weekly assessments, so they had control over how their work was read and interpreted. Students were also offered the chance to include evidence of their performance on weekly assessments during the course to influence their final grade. These implementations lent credibility to the assessments and held students accountable for their work. At the same time, students felt empowered and reassured by their role. These elements help explain some aspects of positive affect and the perception that grading procedures were fair.

Because the course was designed with weekly assessment iterations, students had multiple opportunities to attempt success. The total number of items from the weekly assessments was the same as the total number students were given on midterm exams in the original version of the course. Spacing questions throughout the semester allowed multiple attempts at success and multiple opportunities for students to receive pieces of feedback to inform improvement. If students did not perform well on a weekly assessment, they had the reassurance that they would have another chance to try new methods to attempt success in the upcoming week. The format of the assessments did not vary from week to week; instead, they represented a consistent voice reiterating the importance of data interpretation. This uniformity helped students recognize their improvement during the course from week to week, which provided them with the sense that they were succeeding.

In year 2, efforts had been made to elevate student motivation by readjusting their self-generated scores with one given by the TA. The thinking was that performance might be improved by informing students more realistically of their status relative to a potential grade. This attempt appeared successful since improvements in student performance and self-scoring accuracy were observed (Figures 2C and 3). However, student comments in class and responses on the affective survey (see items 35 and 38 in Table 4) suggested that we were misguided in this approach; it appeared that student attention was focused more on assigning the correct grade than on the more productive elements of feedback. Moreover, some of the survey responses may have also reflected frustrations or disagreements with the TA's perception of their answers to assessment items. Thus, we predicted that de-emphasizing the letter grades combined with increased attention to formative feedback would improve student attitudes while retaining the enhanced performance. Consequently, we abandoned the readjustment of self-assigned scores and focused greater attention on discussion and feedback at the end of each weekly assessment during year 3.

We were gratified to discover that not only were attitudes better during year 3, but performance was increased by an additional increment compared with year 2, and self-scoring accuracy was retained (Figures 2 and 3, and Table 4). Some of this success surely reflects instructor experience and comfort with the course structure. Nevertheless, much of the improvement in year 3 may also be attributed to the quality of feedback students received with each Friday assessment (see item 40 in Table 4). This feedback was immediate and multidimensional; it included personal, peer, and instructor components. On completion of the assessment, students were first invited to spend 5 to 10 min sharing their written responses with one or more nearby partners in the classroom. The instructor then projected to the entire class a set of expert responses to the items with a brief explanation of each. This was followed by an additional 10 min of interaction among students in small groups as they compared their responses with those of the expert. This discussion was animated and noisy with frequent requests for input from the instructor or TA. Students were encouraged to help each other discover ways in which their responses could have been improved and how they might prepare better for the next assessment. Students were then offered the opportunity to remain in the classroom for an additional hour of discussion of the problems. This additional hour consisted of both small group and whole class conversations. Some of these discussions included sharing of individual responses with the entire class and generating a discussion of the merits of those responses. The instructor periodically provided samples of responses at different levels of quality, and students were given practice in evaluating and ranking those responses. The instructor then modeled his evaluation of the

samples. Throughout the various exercises used to provide feedback, the focus was always twofold, helping students obtain a realistic sense of their own performance and teaching them to make decisions regarding how they might improve.

Weekly observations of the class by the instructor supported the idea that the extra hour of feedback offered after assessments was key. Even though attendance was optional and the session was held late in the afternoon on Fridays, most of the students remained and actively participated. The level of engagement contrasted that observed in previous years when the instructor attempted to supplement feedback provided immediately after assessments (commonly confined to ~10 min) by continuing the discussion on the following Monday. Interestingly, students appeared unresponsive and disinterested in further discussion of assessments from the previous week. This observation reinforced the thought that both the adequacy as well as the timeliness of the feedback were essential. Because the extra feedback hour during year 3 occurred immediately after an assessment, students were highly engaged with the material and receptive to learning. We conclude that one of the most important modifications that can be made to any classroom is to engineer situations such as this where students have an opportunity to try out their learning followed by clear and timely feedback.

Notwithstanding all the changes implemented throughout the three years, the reported time students spent outside of class remained constant. Students spent or believed they spent about 2 h less each week than that spent in the original version of the course. To the extent that this time difference was real, an increase in student efficiency is one possible explanation. Alternatively, the reported time difference could be perceptual. Perhaps the time that students reported reflected layout of the course or was affected by direction the instructor gave about how much time students should be spending outside of class. Because performance on data analysis tasks improved despite this decreased time spent, we believe this change to be a positive one.

Faculty attempting to implement this teaching method may be concerned about the potential for grade inflation. In fact, if a criterion-based grading system is used and student performance improves, one would expect an elevation in grades. Nevertheless, this system did not appear to inflate grades inappropriately. For example, in the third year, when performance was the highest, the course grade point average was 3.19 $\pm$ 0.7 (SD). Moreover, the process of empowering students to have input into their course grade had only a modest effect on class grades. In year 3, only 22% of the class successfully justified a higher grade than what was achieved on the final exam. The average grade increment among those students was 0.35 $\pm$ 0.07 (SD) grade point units.

In summary, this study has demonstrated three important lessons. First, it has reinforced and corroborated insights promoted by educational theorists and researchers: frequent, nonthreatening formative assessment is a valuable tool for instructors and students (Butler and Nissan, 1986; Black and Wiliam, 1998; Huba and Freed, 1999; Klionsky, 2001). Second, it has shown that these three aspects of assessment can be achieved in the context of developing higher-order thinking skills in the science classroom. Finally, it has emphasized that implementation of pedagogical reform in a course requires two critical elements analogous to the process described here for student learning: formative feedback and iteration. As described above, the strong student performance and positive attitudes did not appear instantaneously upon adoption of the formative format. Without careful attention to performance and affective survey results, the instructors would not have made those decisions that led to success. Although occasional misinterpretation of the data from these evaluations can result in detours such as occurred for us during year 2, the process is self-correcting if applied consistently. Hence, attention to the data gathered during year 2 generated the ultimate success observed in year 3. Regardless of whether the specific format described in this report seems applicable to other courses or worthy of consideration, the process of course improvement illustrated is imperative. The most important message we could communicate is that all instructors should be actively engaged in systematic evaluation and responsive decision-making in their courses.

ACKNOWLEDGMENTS

The contents of this article were developed under grants from the U.S. Department of Education (P116B980586 and P116B041238). However, those contents do not necessarily represent the policy of the Department of Education, and you should not assume endorsement by the Federal Government.

REFERENCES

Benware, C., and Deci, E. (1984). Quality of learning with an active versus passive motivational set. Am. Edu. Res. J. *21*, 755–765.

Beyer, B. K. (2001). Teaching Thinking Skills—Defining the Problem, 3rd ed. ed. A. L. Costa, Alexandria, VA: Association for Supervision and Curriculum Development, 35–40.

Black, P., and Wiliam, D. (1998). Inside the Black Box: Raising Standards through Classroom Assessment. Phi-Delta-Kappan *80*, 139–147.

Bonwell, C. C., and Eison, J. A. (1991). Active Learning: Creating Excitement in the Classroom. Washington, DC: George Washington University, School of Education and Human Development, Report nr 1. ED340272.

Burrowes, P. A. (2003). A student-centered approach to teaching general biology that really works: Lord's constructivist model put to a test. Am. Biol. Teacher *65*, 491–502.

Butler, R., and Nissan, M. (1986). Effects of no feedback, task-related commands, and grades on intrinsic motivation and performance. J. Edu. Psychol. *78*, 210–216.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. Rev. Edu. Res. *58*, 438–481.

Fink, L. D. (2003). Creating Significant Learning Experiences, San Francisco: Jossey-Bass.

Fitzpatrick, J. L. (2004). Exemplars as case studies: reflections on the links between theory, practice, and context. Am. J. Eval. *25*, 541–559.

Glover, J. A. (1989). The testing phenomenon—not gone but nearly forgotten. J. Educ. Psychol. *81*, 392–399.

Hay, L. (2001). Thinking Skills for the Information Age, 3rd ed., ed. A. L. Costa, Alexandria, VA: Association for Supervision and Curriculum Development, 7–10.

Huba, M. E., and Freed, J. E. (1999). Learner-Centered Assessment on College Campuses, Needham Heights, MA: Allyn and Bacon.

Hughes, B., Sullivan, H., and Mosley, M. L. (1985). External evaluation, task difficulty, and continuing motivation. J. Educ. Res. *78*, 210–215.

Kitchen, E., Bell, J. D., Reeve, S., Sudweeks, R. R., and Bradshaw, W. S. (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.

Klionsky, D. J. (2001). Constructing knowledge in the lecture hall. J. College Sci. Teach. *31*, 246–251.

Kohn, A. (2004). What Does It Mean to Be Well Educated?: And More Essays on Standards, Grading, and Other Follies, Boston, MA: Beacon Press.

Lodish, H., Berk, A., Matsudaira, P., Karp, G., Krieger, M., Scott, M. P., Zipursky, S. L., and Darnell, J. (2003). Molecular and Cellular Biology, 5th ed., New York: W. H. Freeman and Company.

Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., and Darnell, J. (2000). Mol. Cell. Biol. 4th ed., New York: W. H. Freeman and Company.

McConnell, D. A., Steer, D. N., and Owens, K. D. (2003). Assessment and active learning strategies for introductory geology courses. J. Geosci. Edu. *51*, 205–216.

National Research Council (1996). National Science Education Standards: Observe, Interact, Change, Learn, Washington DC: National Academy Press.

National Research Council (2000). How People Learn: Brain, Mind, Experience and School, Washington, DC: National Academy Press.

National Research Council (2001). Classroom Assessment and the National Science Education Standards, Washington, DC: National Academy Press.

O'Sullivan, D. W., and Copper, C. L. (2003). Evaluating active learning: a new initiative for a general chemistry curriculum. J. College Sci. Teach. *32*, 448–452.

Rifkin, T., and Beorgakakos, J. H. (1996). Science reasoning ability of community college students. ED393505.

Stiggins, R. J., and Conklin, N. F. (1992). Teacher's Hands: Investigating the Practices of Classroom Assessment, New York: State University of New York Press.

Weimer, M. G. (2002). Learner-Centered Teaching: Five Key Changes to Practice, San Francisco: Jossey-Bass.