CBE—Life Sciences Education
Vol. 8, 131–139, Summer 2009

Article

A Statistical Analysis of Student Questions in a Cell Biology Laboratory

Elena L. Keeling,* Kelly M. Polacek,* and Ella L. Ingram†

*Department of Biological Sciences, California Polytechnic State University, San Luis Obispo, CA 93407; and †Applied Biology and Biomedical Engineering, Rose-Hulman Institute of Technology, Terre Haute, IN 47803

Submitted September 11, 2008; Accepted February 2, 2009
Monitoring Editor: Laura Mays Hoopes

Asking questions is an essential component of the practice of science, but question-asking skills are often underemphasized in science education. In this study, we examined questions written by students as they prepared for laboratory exercises in a senior-level cell biology class. Our goals were to discover 1) what types of questions students asked about laboratory activities, 2) whether the types or quality of questions changed over time, and 3) whether the quality of questions or degree of improvement was related to academic performance. We found a majority of questions were about laboratory outcomes or seeking additional descriptive information about organisms or processes to be studied. Few questions earned the highest possible ranking, which required demonstration of extended thought, integration of information, and/or hypotheses and future experiments, although a majority of students asked such a question at least once. We found no correlation between types of student questions or improvement in questions and final grades. Only a small improvement in overall question quality was seen despite considerable practice at writing questions about science. Our results suggest that improving students' ability to generate higher-order questions may require specific pedagogical intervention.

INTRODUCTION

> "Fundamental questions are guideposts; they stimulate people. One of the most creative qualities a research scientist can have is the ability to ask the right questions."
>
> (Gross, 2004 Nobel laureate in physics[1])

Questioning is a fundamental skill expected of scientists and scientifically literate citizens. This expectation is reflected in the National Science Education Standards, which describe questioning at every grade level in the "Content Standards for Science as Inquiry" (National Research Council, 1996). Although asking questions is an essential part of doing science, questions are often underemphasized in how we practice science education. In traditional classrooms, the majority of students spend little time asking questions, and it is rare for students to receive feedback on the scientific quality of their questions. Although scientists clearly appreciate the value of questions, students in science classes may not. A low frequency of oral student questions has been observed at all grade levels and across all subjects (Dillon, 1988).

In addition to the critical link between questions and the process of science, student questions can be valuable for a variety of pedagogical reasons. On a basic level, assigning students written questions on readings outside of class requires them to actually do the reading and come to class more prepared (Marbach-Ad and Sokolove, 2000a; Polacek and Keeling, 2005). Writing questions may help students focus their attention on a text or lecture topic, increasing their understanding. Rosenshine *et al*. (1996) analyzed 26 studies in which students were taught to generate questions as they read texts. They found an increase in comprehension on subsequent exams. Questioning may facilitate development of analytical and critical-thinking skills. For example, Zoller (1987) emphasizes question-asking as an essential skill for solving problems. Student questions, both written and oral, reveal misunderstandings, confusion, misconceptions, and interests.

[1] As quoted in Siegfried (2005).

DOI: 10.1187/cbe.08–09–0054
Address correspondence to: Elena L. Keeling (ekeeling@calpoly.edu).

A variety of creative strategies have been used in attempts to elicit student-generated written questions about science. Students have been asked to write questions on science texts in class (Shodell, 1995; Costa *et al.*, 2000) as well as to write their own questions and then answer them on exams (Zoller, 1987). Student questions may be in response to textbook readings (Marbach-Ad and Sokolove, 2000a,b), individual observations (Marbach-Ad and Claassen, 2001), case studies (Dori and Herscovitz, 1999), or research papers (Brill and Yarden, 2003). Those studies were carried out with high school students or freshmen in college. We have been unable to find reports analyzing student questions in upper-division college science classes. If exposure to advanced courses improves students' ability to think in a more sophisticated manner about science, we would expect this to be reflected in a higher quality of student questions at the senior undergraduate level.

Questions were categorized differently in each of the studies cited above, but with a shared emphasis on higher-order or more scientific questions. These questions were identified as those involving synthesis of prior knowledge or proposal of a hypothesis (Marbach-Ad and Sokolove, 2000a,b); "deep reasoning questions" (Costa *et al.*, 2000); questions about causality (Brill and Yarden, 2003); and complex questions involving application, analysis, or evaluation (Dori and Herscovitz, 1999). Improvements in the quality of high school students' questions were found after classroom discussion of science-related case studies (Dori and Herscovitz, 1999) or an original scientific research article (Brill and Yarden, 2003). At the college level, Marbach-Ad and Sokolove (2000a,b) found that explaining the categorization and providing examples increased the number of higher-level questions; this effect was stronger in an active-learning class that emphasized in-class discussion and student questions as well as written homework questions.

The laboratory setting would seem to present an ideal opportunity for stimulating high-quality student questions. Laboratory courses are designed to involve students in the process of doing science, expose them to experimental techniques, stimulate careful observation, and require thought about data to draw conclusions. Each of these facets lends itself to instigating student questions regarding the science being done. Many laboratory activities revolve around students attempting to answer a scientific question; even if the questions addressed are stated in the laboratory manual rather than generated by students, they should serve as models. Thus far, there has been little analysis of student questioning abilities in the laboratory. One study (Marbach-Ad and Claassen, 2001) examined student questions in a college introductory biology laboratory course; initial questions were in response to a biology cartoon and final questions were based on observations and intended to be research questions the students would have the option of investigating in lab.

In this study, we examined written student questions in reference to laboratory exercises in a senior-level cell biology class. The laboratory exercises included background material, detailed protocols, types of data to be recorded, and guidelines for analysis. As such, reading this material before class provided an opportunity for students to think about the scientific process and specific biological topics. Question-writing was part of a prelab assignment, so that students wrote questions nearly every week. This allowed us to determine whether question-writing ability improved with practice. Our goals were to discover 1) what types of questions students ask about biology laboratory activities; 2) whether the types or level of the questions changed over time; and 3) and whether the quality of questions or degree of improvement was related to academic performance in the class.

METHODS

The study was carried out in a senior-level cell biology class taught during a 10-wk quarter. There were 40 students in the class, but two declined to have their work included in the study; permission was requested at the end of the quarter so as not to bias the questions written. The study was approved by the Cal Poly Human Subjects Committee. Of the 38 students in the study, 95% were biological sciences, microbiology, or biochemistry majors; 87% were seniors and 11% juniors; and 55% were women. The mean course grade was 79.8 $\pm$ 9.9% (1 SD, used throughout this report), whereas the mean lab grade was 83.5 $\pm$ 7.8%. The lab grade was 25% of the final course grade.

As a part of prelab assignments, students wrote at least three specific, concrete questions that arose as they read the laboratory exercise and thought about the upcoming experiments. Students were directed that at least one question should be a question addressed in or resolved by the laboratory activity. The instructions encouraged students to think about connections and possible applications while discouraging questions about the definition of a word or why a particular reagent was being used. Questions were submitted at the beginning of each of eight laboratory sessions. For simply submitting their three questions, students earned full credit with the following exceptions: questions that did not follow the guidelines (e.g., asked for the definition of a word) did not receive credit, whereas those demonstrating unusual amounts of thought or creativity earned extra credit. Brief marginal comments were written for feedback, discouraging lower-level questions and attempting to prompt greater clarification and deeper thought. The question assignments accounted for ~5% of the total lab grade. The laboratory activities used in the study are briefly described in Table 1. Other aspects of the laboratory course and questions are described in Polacek and Keeling (2005).

To classify questions, we began by examining prelab questions from students in a previous laboratory section and identifying types asked. Categories were further refined based on examination of published literature (West and Pearson, 1994; Shodell, 1995; Watts *et al.*, 1997; Dori and Herscovitz, 1999; Costa *et al.*, 2000; Marbach-Ad and Sokolove, 2000a; Brill and Yarden, 2003). Ultimately, we modified the scheme developed by Marbach-Ad and Sokolove (2000a,b) to make it appropriate for categorizing questions generated in preparation for a specific experimental procedure. Our six categories are as follows, ranked from low level to high level.

Category 0. Questions that do not make logical sense, are based on a fundamental misunderstanding, are too general to be meaningful, or are not relevant. For example, "What will this tell us about how structure is related to function?," "Is this the same yeast used to make beer?," and "Who was Robert Hill?"

Category 1. Questions about a simple definition, expected knowledge, or that are clearly answered in the reading material. Examples include "What is SDS?" and "Why do we use colchicine?"

Category 2. Questions that should be answered directly by observation in the laboratory session. For example, "Which cells will move faster?" and "Which inhibitor will have the greatest effect on the yeast?"

Table 1. Lab titles and description of experimental techniques used and biological phenomena observed

Lab title	Experimental techniques; biological phenomena
Lab 1: Microscopy & Cell Structure and Function	Light microscopy, cell staining; observation of protozoa and plant cells, inhibition of actin
Lab 2: Cell Fractionation & Photosynthesis	Differential centrifugation, spectrophotometry; electron transport, photosynthetic pigments
Lab 3: In Vitro Protein Interactions	Protein isolation, gel electrophoresis; GST fusion proteins, SNARE protein complexes
Lab 5: Cytoskeleton: Flagellar Regeneration	Microscopic measurements; microtubule polymerization, flagellar motility
Lab 6: Cytoskeletal Organization & Fluorescence Microscopy	Fluorescence microscopy; organization of actin and microtubules
Lab 7: Yeast Cell Cycle	Fluorescence and light microscopy; temperature-sensitive mutants, cell cycle arrest
Lab 9: Signaling: Yeast Mating Factor Response	Spectrophotometry, light microscopy; β-galactosidase reporter activity, cell cycle arrest, signaling pathway
Lab 10: Blood Cell Culture & Adhesion	Light microscopy, neutrophil enrichment, cell staining; adhesion, extracellular matrix, phagocytosis

Notes: Lab 3 requires 2 weeks so there is no separate Lab 4. Lab 8 is an independent student-designed experiment. The lab manual was written by Elena L. Keeling and Michael Black, Biological Sciences, California Polytechnic State University.

Category 3. Questions going beyond what will be seen in lab but not addressing mechanisms or explicitly integrating information. They include questions about evolution and purpose, questions seeking additional descriptive information about phenomena, and those making simple connections to other knowledge or to applications. Examples include "How did photosynthesis evolve?," "Why does the cell cycle arrest in response to a pheromone?," "What is the pH of the natural environment of *Chlamydomonas*?," and "How is the drug cytocholasin used in humans?"

Category 4. Questions asking about mechanism, or how things work at a cellular or molecular level. For example, "Why does methyl green have different affinity for different things?" and "How is it that tubulin heterodimers are transported to the distal end of the growing flagellum?"

Category 5. Questions that reveal extended thought and integration of information; often include a prediction or a hypothesis about a possible follow-up experiment. These questions are often preceded by a statement of perceived paradox or something puzzling. For example, "Why is it important to keep the samples on ice at all times when these cells and chloroplasts are not ice cold in nature?," "How does Protoslo slow the protozoa down; does it affect the cell internally or does it change the environment, such as the viscosity of the solution?," and "Dyes are normally large molecules and one would think that these large molecules would have difficulty passing through the membrane via diffusion alone. Do reagents have to be functionalized with receptors or signaling groups to permeate the cell membrane and dye organelles?"

Questions in both categories 3 and 4 contain elements important for science; the explicit focus on mechanism in category 4 was considered more revealing of experimental scientific thinking. Category 5 contains the highest quality scientific questions. A small degree of confusion or misunderstanding was accepted in higher-ranked questions if it was a reasonable misunderstanding and the question showed evidence of analysis or synthesis.

Questions were transcribed into an Excel spreadsheet (Microsoft, Redmond, WA) and given a letter code corresponding to each student so that they were anonymous when categorized. We used 270 questions spanning three different lab exercises from a previous class to refine the categories, practice rating questions, and clarify our categorization guidelines. For this study, questions were assigned to categories independently by Polacek and Keeling. Initial categorizations yielded between 56 and 72% agreement. Discussion of the remaining questions led to 96–100% agreement. Any questions that remained controversial were given the lower of the two possible rankings to be conservative about attributing higher-order thought.

DATA ANALYSIS

We describe our analyses here to illustrate various mechanisms for exploring categorical data common to many educational research studies. During initial analysis of the data, we arbitrarily assumed a single unit of question type separating each category; this simplification has been successfully used by other investigators (e.g., Marbach-Ad and Sokolove, 2000) to reveal general trends. To assign a score to the questions for each student each week, we calculated a quality score in which the number of questions in each category was multiplied by the category value and scaled. For example, a student who wrote one category 2 and two category 3 questions earned 8 points [$(1 \times 2) + (2 \times 3) = 8$]. This point value was then converted to a percentage, with a maximum score being 15 points or 5 points $\times$ the number of questions asked if more than three. This strategy allowed us to obtain a single value representing the quality of each student's questions for each assignment. We calculated two normalized gain scores; these scores allow us to evaluate increases in student questioning ability relative to their starting point (Hake, 1998). This scaling is informative because students' initial question ability was quite varied (as described in *Results*). We calculated the baseline gain score by subtracting the Lab 1 quality score from the Lab 10 quality score, then normalizing by dividing by the maximum number of points that could have been gained (i.e., maximum quality score [i.e., 100%] $-$ Lab 1 quality score). The maximum gain score was calculated in the same manner, except that the highest-quality score, regardless of lab, was used in place of the Lab 10 quality score. Because only 11 of 38 cases had identical baseline and maximum gain scores, we felt these two separate analyses were appropriate.

All response variables (e.g., quality score, course grade, and number of questions) were tested for normality using the Anderson–Darling test. This test evaluates the distribution of data relative to a normal distribution, as do other normality tests, but is generally understood to manage data in the tails of the distribution better than other tests (D'Agostino and Stephens, 1986; although we note that different tests of normality almost always give identical results, as for our data, except for very highly skewed distributions). Lab grade and total number of questions asked by individual students were not normally distributed. The nonnormality of the total questions variable was expected, as students were instructed to submit three questions per lab (resulting in 24 questions). The nonnormality of the lab grade variable was strongly influenced by two students receiving scores in the 60s, but we retained these data because our sample size was small and further testing revealed no meaningful

Table 2. Scoring and distribution of questions among categories for the eight labs analyzed

Category[a]	Type of question	Lab							
		1	2	3	5	6	7	9	10
0	Noninformative	1.6[a]	10.3	13.7	1.7	2.6	3.5	10.0	1.8
1	Definitional	4.9	11.1	15.4	3.4	1.7	4.4	5.5	5.4
2	Observational	54.5	21.8	8.5	33.1	32.8	34.5	26.4	24.3
3	Connection or application	28.5	30.8	43.6	37.3	42.2	41.6	43.6	45.9
4	Mechanism	6.5	15.4	12.8	16.9	12.1	9.7	6.4	17.1
5	Hypothesis or prediction	4.1	7.7	6.0	7.6	8.6	6.2	8.2	5.4
High score[b]		4	10	3	9	7	4	5	9
Mean score[c]		49.7	51.2	49.2	58.0	56.9	54.0	51.8	57.6
n[d]		38	36	37	37	37	36	35	36

[a] Values listed for categories of questions are percentage of questions submitted fitting the category.

[b] Indicates number of students whose highest question score occurred in that week.

[c] Refers to the class mean question score for each lab (scaling from 0 to 100; essentially mean percentage).

[d] Number of students submitting questions for that lab.

influence of these scores on the results. We evaluated the results of the nonparametric comparison of independent samples using the Mann–Whitney *U* test (which does not assume normally distributed data, instead comparing ranks of values rather than the values themselves) and the parametric equivalent Student's *t* test of comparison of independent samples for these variables (comparing for example, mean number of questions asked among biology majors to nonbiology majors; Zar, 1999); the outcomes were identical, so only those results from parametric tests are reported, because these tests are more familiar.

We compared the distribution of questions among the categories between labs occurring later in the term and the first two labs. We chose Lab 1 and Lab 2 as our baselines of comparison because the questions submitted for these labs represent the initial questioning ability of the students. This comparison was accomplished using a χ^2 test, also known as a contingency test. A common tool from genetics analysis, the χ^2 test summarizes differences between some known categorization (the expected) and the actual results received (the observed values) by comparing frequencies of occurrence (Zar, 1999). In our case, the expected values are derived from either the Lab 1 question distribution or the Lab 2 question distribution. This test allowed us to evaluate global changes in the questions of the class as a whole, rather than individual students' changes.

To examine whether questioning ability was related to achievement, we performed regression analysis, using our various measures of questioning ability individually as potential predictors of either final course grade or lab grade. We used the same statistical tool to examine change in the class's questioning ability over time, with week in the course as a predictor of proportion of questions attributed to each category. This same issue was examined for each individual student using the Pearson r, with time and question score as variables of interest.

Our criterion for statistical significance was $p < 0.05$. When we tested numerous comparisons simultaneously, as in Table 4, we applied a Bonferroni correction to account for the known observation of an increased likelihood of detecting a statistically significant outcome based purely on the high number of tests being performed (basically, an increased probability of a type I error; Bland and Altman, 1995). This correction was accomplished by resetting the criterion for statistical significance at 0.05 divided by number of simultaneous tests performed. Results were then considered statistically significant if the resulting *p* value was smaller than the new, more conservative threshold. The correction was applied table-wide. These corrections are noted in *Results*.

RESULTS

In total, 924 questions were analyzed, ranging from 110 to 123 per lab. The number of total questions asked per student ranged from 15 to 33, with a mean of 24 ± 3. Seventy-six percent of the students submitted all eight assignments, whereas 18% missed one assignment; one student missed two assignments, and one student missed three assignments. With nonresponses removed, the overall mean qual-

Table 3. Distribution of students asking higher-order questions at any time during the class[a]

No. of category 3, 4, or 5 questions	% of students	No. of category 4 or 5 questions	% of students	No. of category 5 questions	% of students
0	0	0	7.9	0	39.5
1–5	2.6	1–5	57.9	1	23.7
6–10	15.8	6–10	28.9	2	15.8
11–15	39.5	11–15	5.3	3–7	21.1
16–20	36.8				
21–25	5.3				
Any	100	Any	92.1	Any	60.5

[a] Mean number of questions asked per student 24 ± 3.

Table 4. χ^2 results comparing distribution of questions in subsequent labs to early student performances

Lab	Lab used as expected distribution	
	1	2
3	165.7**[a]	19.6*
5	31.5**	16.8*[b]
6	27.8**	21.3**
7	18.4*	18.4*
9	70.0**	13.7*[b]
10	45.3**	18.3*

* $p < 0.05$, ** $p < 0.001$.

[a] For example, χ^2 value resulting from the contingency test of data presented in the Lab 1 column to the Lab 3 column of Table 2.

[b] Not significant following Bonferroni correction.

ity score for the question assignment through the term was 54 ± 9%. For the term as a whole, the majority of questions were in categories 2 and 3; very few questions were entirely off-topic or irrelevant (Table 2).

All students wrote some questions in categories 3, 4, or 5; approximately four-fifths of the class (82%) wrote more than 10 (Table 3). When considered as a percentage of the total number of questions written by each individual, students wrote a mean of 59% questions at this level, with a range of 22–84%. Focusing on the two highest categories, 92% of students wrote at least one category 4 or 5 question and 34% wrote more than five questions. Finally, 61% of students wrote at least one category 5 question and 21% wrote three or more questions. Four students (11% of the class) wrote a category 5 question for at least half of the assignments.

We analyzed the questions from each lab relative to each other, to determine whether questioning ability changed when explicit practice in formulating questions was provided. Student questioning ability improved somewhat over the term, based on several measures. First, when comparing the quality scores for Labs 1 and 10, 64% of the class demonstrated an increase in questioning ability, 8% showed no change in questioning, and 28% showed a decrease (n = 36 because two students did not turn in one of these assignments). Second, the distribution of questions among the categories for Labs 1 and 2 served as two baselines for comparison. We found that the distribution of questions in all subsequent labs differed significantly from these baselines (Table 4). After applying a Bonferroni correction, 10 of these 12 comparisons were still statistically significant. A different distribution from baseline does not by itself imply an improvement; the large difference between Lab 3 and Lab 1 reflects increases in categories 0 and 1 as well as categories 4 and 5, and an unusual decrease in category 2. However, in comparison with Lab 1, all subsequent labs have fewer questions in category 2 and more questions in category 5. In comparison with Lab 2, all subsequent labs have more questions in category 3. Third, both the baseline gain score and the maximum gain score were statistically significantly different from zero (baseline gain: mean 0.11 ± 0.32, $z = 2.01$, $p = 0.022$, df = 36; maximum gain: mean 0.40 ± 0.31, $z = 8.05$, $p < 0.001$, df = 38; both tests were one-tailed consistent with our expectation of positive change in questioning over the term). Student questioning ability was better later in the term than the starting point of the first questioning assignment.

We extended our understanding of whether this increase in questioning ability was linear (students getting gradually better through the term) or idiosyncratic (best performances occurring throughout the term) by determining when the best performances occurred in the term. These best performances match to the maximum gain score described above. Most students had their best performances in Labs 2, 5, or 10 (Table 2). Furthermore, we expected an increase in questioning ability over the term to be demonstrated by an increase in the proportion of higher-order questions asked. In a regression analysis, there was a trend toward a decrease in questions in categories 0, 1, and 2 (negative coefficients b), whereas there was a trend toward an increase in other categories over time; however, none of these results were statistically significant. When we pooled lower-order questions (categories 0, 1, and 2) and higher-order questions, the same outcome occurred—neither group showed a statistically significant change over time, although lower-order questions decreased in frequency from 60% of questions to 31% of questions by Lab 10. We noted that after the first questioning assignment, at least half of all questions submitted fell into the higher-order question categories. On a whole-class basis, the questioning ability did not increase linearly through the term. When analyzed for each student individually, we found a slight bias toward increased questioning ability through the term, based on two results. By determining the correlation between time and question score for each student, we found half of the students demonstrated little correlation between time and questioning performance ($r < 0.3$). We found a positive correlation greater than $r = 0.3$ between these measures for 13 students (three of these correlation values were statistically different from zero); of the six students with negative correlations less than $r = -0.3$, none were statistically significant. In addition, we found statistically significantly higher questioning scores in Labs 5 and 10 compared with Lab 1 (Lab 1 compared with Lab 5: $t = -2.90$, $p = 0.003$, n = 37; Lab 1 compared with Lab 10: $t = -2.84$, $p = 0.004$, n = 36; in both cases, paired t tests were performed). The quality of student questions tended to be higher later in the term, but this pattern was not robust.

We found that no aspect of student questioning ability predicted course achievement measured as either the lab grade or as final course grade (students who failed to submit an assignment were excluded for that predictor). None of the individual lab question quality scores, the mean question quality score, nor the two gain scores was a statistically significant predictor in a simple linear regression. This finding included all of the questions submitted by students. When only the three best questions submitted for each lab assignment were used as predictors of lab grade or final course grade, we found that performance on the questioning task for Lab 2 was a weak predictor of final grade ($b = 0.154$, $F = 4.71$, $p = 0.037$, adjusted $R^2 = 9.1\%$, n = 37). However, this result did not pass a Bonferroni correction for the number of simultaneous tests we performed. Questioning ability did not predict achievement in this course.

Table 5. Comparison of dependent variables between dichotomous demographic characters

Demographic status (n)	Final course grade			Final lab grade			Total questions asked			Mean question score		
	Mean (SD)	$t^{a,b}$	df	Mean (SD)	t	df	Mean (SD)	t	df	Mean (SD)	t	df
Female (21)	81.8 (7.8)	1.38	26	85.2 (7.4)	1.57	33	24.8 (3.0)	1.09	33	53.6 (9.9)	0.42	35
Male (17)	77.2 (11.9)			81.3 (7.9)			23.7 (3.2)			54.8 (7.7)		
Major (30)	81.1 (10.1)	1.88	14	85.1 (6.0)	1.95	8	24.7 (2.7)	1.04	8	52.8 (9.1)	2.27	16
Nonmajor (8)	74.9 (7.8)			77.4 (10.8)			23.0 (4.3)			59.1 (6.2)		
Nonsenior (5)	78.5 (8.1)	−0.38	6	73.7 (12.5)	−1.99	4	22.6 (4.3)	−1.00	4	57.0 (11.2)	0.63	4
Senior (33)	80.0 (10.3)			85.0 (5.7)			24.6 (2.9)			53.7 (8.8)		

[a] All t tests were two-tailed with unequal sample sizes and unequal variance.

[b] No t tests were statistically significant following Bonferroni correction.

We performed various additional analyses to discover whether demographic characteristics influenced questioning or questioning ability. There was no significant difference between male and female students in either final course grade or final lab grade, although in both cases, females earned slightly higher grades than did males (a 4% difference; Table 5). We found no difference in the number of questions asked through the term between genders. Considering the mean quality score for all submitted questions, we also found no difference between genders in the level of questions asked. The sample population contained 30 biology majors and eight nonbiology majors (e.g., agriculture, microbiology). On average, the biology majors earned ~7% higher grades than the nonbiology majors, but these differences were not statistically significantly different. Furthermore, the two groups did not differ in terms of the number of questions asked through the term. The questioning ability of the nonmajors was slightly higher (6%) than the majors, although this comparison is not statistically significant after a Bonferroni correction for multiple simultaneous comparisons. Seniors numerically dominated the sample population, with 33 of the 38 students reporting senior status. Given this large inequality in the sample population, we tentatively analyzed our dependent variables of interest, comparing seniors to all nonseniors. These two groups did not differ significantly for any of our dependent variables. In total, demography had no effect on course achievement or questioning.

DISCUSSION

A crucial issue is how to categorize student questions to evaluate their quality. We hoped to use a categorization scheme developed by others, but we found that none was completely appropriate for questions about laboratory experiments. Our question scheme is similar to that developed by Marbach-Ad and Sokolove (2000a,b), with the greatest difference in category 2. Our students were instructed to write at least one question that would be answered during lab; these observation-based category 2 questions are not included in previously published categorization schema. In contrast, category 2 for Marbach-Ad and Sokolove contains "ethical, moral, philosophical, or sociopolitical questions"; we received very few of these, and they were scored as either category 3 or category 0 depending on their degree of connection to science. Our category 3 is broader than theirs and our category 4 is narrower, restricted to questions addressing mechanisms. Questions about mechanisms and causality are similarly ranked highly in several other categorization schemes (Brill and Yarden, 2003; Costa *et al.*, 2000). Finally, our category 5 is a merger of their categories 5 and 6 containing all of the most thoughtful questions.

As reported by other investigators, we found some questions difficult to categorize definitively. Barden (1995) notes that even classifying questions as lower versus higher level may be complicated because the thought processes required when writing a question depend heavily on context, particularly what information was included in the reading or previous instruction. In addition to immediate context, the level of thought required to write a question can vary depending on prior knowledge. Furthermore, any categorization is an inference that relies on sometimes subtle word choices. In their analysis of questions written by eighth-graders, Arzi and White (1986) found through interviews that students did not always distinguish "what" from "why"; some questions that seemed to be higher-order questions about causes were not intended that way. In some cases students seemed to simply model questions after those they had heard in class without thinking deeply themselves. College students also seem to use why to mean both *how* things occur at a mechanistic level, and *why* things occur in terms of evolution. When students did not provide enough information to allow us to distinguish these two meanings, questions were scored more conservatively.

Our first goal was to determine what kinds of questions students ask when reading material related to science laboratory activities. Guidelines for writing questions were intentionally brief and general to elicit unbiased questions reflecting the range of student thought. Students were specifically directed to write a question that would be answered by the laboratory activity (category 2); this direction was not designed to encourage higher-order questions but rather to focus student attention on the sort of information they would be gathering and encourage them to think about laboratory activities from the perspective of answering questions rather than simply classroom exercises. Given these instructions, we expected at least one-third of the questions would fall into category 2. In fact, for five of the eight

assignments, the percentage was reasonably close to a third (24.3–34.5%; Table 2). Possible reasons for variation in question distribution across different assignments are discussed below.

Most student questions focused on lab outcomes (category 2) or revealed curiosity about the scientific topic (category 3). Whereas category 2 and 3 questions were most common, students were capable of thinking about mechanisms because >90% wrote at least one category 4 question. The majority (60%) wrote at least one category 5 question, demonstrating the ability to engage in higher-order thinking. Two published analyses of student questions using a similar categorization scheme in introductory biology classes reported that without class training or discussion of question quality the mean percentages of students writing the highest-order questions were 11.5, 25, and 27% in different classes (Marbach-Ad and Sokolove, 2000b; Marbach-Ad and Claassen, 2001). After explicit training and class discussion of question quality this increased to 17, 44, and 41%, respectively. A greater number of students writing more scientific questions at the senior level would support the idea that science education improves this skill. However, several factors confound the comparison: the different number of questions written per student, aggregate instead of individual data, and different material used to elicit student questions. It is possible a larger number of introductory biology students would have written higher-order questions given a comparable number of opportunities as our students. Data in the published studies were reported as percentages for each assignment, so it cannot be determined whether the same students wrote category 5 questions multiple times or whether larger numbers of students wrote a category 5 question at some point. In fact, our percentage of category 5 questions for a single assignment (4.1–8.6%) was substantially lower than those classified as category 5 or 6 written by introductory students for most single assignments. The effect of substantially different prompts for student questions is unclear. Finally, it is also possible that different degrees of rigor were used when placing questions in the highest category by different authors. It would be interesting to investigate the differences between freshmen and seniors in a single study minimizing other variables.

From an anonymous survey at the end of the term, many students identified the point of pre-lab questions as helping them prepare for the lab. Some students found this useful; one wrote that "asking questions about the lab invoked more interactive thinking and helped tremendously with understanding the lab and its protocols." Other students found question-writing tedious or a waste of time; one commented "sometimes, it was really hard to come up with questions since the experiment was straightforward." Only four of 38 students (11%) commented on additional benefits of question-writing. Their comments were "Prelab questions were fun, they allowed for more open-ended thinking about variables in the experiment"; "Interesting to think of other possible experiments or results"; "I really liked the questions we could make up that had nothing to do with the questions answered in lab"; and "I honestly noticed myself being able to ask questions in lecture and in other classes a lot more than ever before." These comments indicate that student attitudes toward the assignment varied widely and were likely to impact the types of questions written. The final comment states a perceived increase in the student's own question-asking and on the transfer of this skill to other settings; it would be interesting to investigate whether this phenomenon was found by other students.

As a whole, the class became slightly better at generating the scientifically testable questions we consider necessary to the practice of science. When considering Lab 1 as a pretest, gain scores were positive though not large, there was a slight increase in categories 3, 4, and 5, and 64% of the class showed an increase in question quality. Improvements in the scientific quality of student questions may have been influenced by such factors as additional readings in a textbook that emphasized the role of specific experiments; stronger understanding of the principles and approaches of cell biology; increased exposure to questions about cell biology and about experimental approaches posed by the lecture instructor; and additional practice at writing laboratory discussions that required analysis of data, proposing hypotheses, and drawing conclusions.

However, this increase in questioning ability was neither linear nor dramatic. Our expectation was that students would gradually become better at writing questions with practice. Instead, we found that best performances were scattered throughout the term. There are a variety of possible explanations for this outcome. Marbach-Ad and Sokolove (2000a) found that students ask more sophisticated questions about topics they understand better. Our students wrote most of their highest-ranked questions for Labs 2, 5, and 10, for which the experimental approaches were relatively straightforward. Conversely, the most unusual distribution of questions occurred for Lab 3, which had the highest percentages of category 0 and 1 questions (29.1% combined) and by far the lowest percentage of category 2 questions (8.5%). During Lab 3, students performed a complex experiment involving expression of recombinant proteins, affinity techniques, and gel electrophoresis; it is not surprising they submitted more lower-order questions about this lab. In Lab 9, students were introduced to reporter gene assays and specific activity calculations; whereas the percentage of category 5 questions was the second highest for this lab (8.2%), there was also a jump in category 0 questions to 10%, suggesting confusion in a larger fraction of the class.

A variety of other factors could contribute to fluctuating performance throughout the course. The introductory material for each laboratory exercise varied considerably, providing differing amounts of information to trigger connections or hypotheses. Student interest in the topics also varied, with students generally showing the most excitement about Lab 6, in which they did fluorescence microscopy, and Lab 10, in which they worked with their own blood. Outside opportunities to learn more about relevant topics varied. For example, students had not yet learned about SNARE proteins before they wrote questions for Lab 3, but they had learned about actin and microtubules before Lab 6. External events also may have affected the amount of time students devoted to the assignment, based on the timing of exams or other assignments in this and other classes. For example, Lab 7 questions were submitted at a time when students were developing plans for an independent experiment as well as taking midterm exams in other classes.

The lack of a dramatic improvement in the scientific quality of student questions despite considerable practice supports the conclusions of other studies that explicit instruction about what constitutes higher-order questions is important for improvement. When students were instructed to write questions that might serve as the basis for actual experiments, knew they would be graded on question quality, and had received explicit instruction in scientific question quality, they asked more questions involving synthesis and proposing hypotheses (41 vs. 28%; Marbach-Ad and Claassen, 2001). When students engaged in small-group discussions including ranking questions, they asked higher-quality questions than those without this experience (30 vs. 13%; Marbach-Ad and Sokolove, 2000b). In the same study, presentation of the categorization scheme with examples resulted in even larger numbers of more thoughtful questions (Marbach-Ad and Sokolove, 2000b). Interestingly, one study reported an increase in higher-level questions without explicit guidance about question quality or classification (Brill and Yarden, 2003); this study population was high school students reading modified research papers in developmental biology. Several attributes of research papers seem similar to those found in the laboratory manual used in this study: exposure to experimental approaches and procedures, an emphasis on experiments being done to answer a scientific question, data analysis. Although these are different student populations, it suggests that reading primary literature may contribute additional benefits to how students think about science.

Questioning ability measured through the course was not a predictor of final course performance. This was true even for the best question-askers; the four students who asked a category 5 question for half or more of the labs earned final grades of "A−," "B−," "C," and "C." Several factors are likely to contribute. Completing the questioning assignment fulfilled the course obligation, i.e., there was no incentive to ask higher-level questions. Additionally, the questioning assignment was a small part of the final course grade. Students may have felt that question-asking was not relevant to other aspects of the course; indeed, the lecture exams (which constitute 75% of the final course grade) did not include an opportunity to demonstrate questioning skills. It does suggest that academic ability was not an important factor in the quality of questions asked. These results do not match those seen for high school students responding to case studies, where students who ranked higher academically showed a greater increase in the number and complexity of questions asked (Dori and Herscovitz, 1999). The reasons for this difference are unknown; two possible factors are the wider range of academic abilities in high school and differences in the basis for final grades. There were also no demographic factors affecting questioning ability in our relatively homogeneous group of students. Other analyses of written questions also have found no gender differences, although under some circumstances there are differences in the number of oral questions asked by male and female students (Pearson and West, 1991).

In our study, questioning was performed by students before coming to class. We provided limited instruction to encourage students to ask whatever questions seemed meaningful to them. As students performed laboratory exercises, we encouraged them to make notes of questions as they arose. We required students to submit follow-up questions at the end of each experiment to emphasize that science is not a discrete set of laboratory exercises that are completed once the lab is finished, but rather in science, questions beget more questions in a positive and progressive way. Our goal was to create an atmosphere of question-asking and hypothesizing that emulates the scientific method.

Our hope was that the practice of asking questions would increase students' question-asking ability. In future studies, however, we will provide explicit instruction on how to formulate meaningful scientific questions. Rather than emphasizing practice by requiring multiple questions to be written each week, it may be more effective to focus more attention on fewer question assignments. Students could be given multiple opportunities for feedback on the merit of their questions; they could analyze each other's questions and revise their own questions in light of this feedback. Several other authors have described specific interventions such as ranking questions in small groups (Marbach-Ad and Sokolove, 2000b) and reflecting on which questions were most informative in helping to solve particular problems (Ingram *et al.*, 2004). It also may be useful to emphasize questions written after completion of laboratory activities and analysis of data; students may exhibit more higher-order thinking after extended interaction with the scientific topic. In fact, this may be a factor in the more dramatic improvement seen by other investigators after students read primary literature (Brill and Yarden, 2003). Another possible change would be an adjustment of course grades to increase the points associated with written questions and provide more incentive for higher-order questions.

Our results suggest that a variety of factors may influence students' ability to write meaningful, higher-order, scientifically testable questions. Repeated practice at writing questions in a class where questions were emphasized resulted in a small degree of improvement in the depth of student questions. This suggests that "learning by doing" is not enough to yield dramatic improvements in this particular cognitive skill, and that more explicit guidance and discussion may be required. Continued investigation is needed to identify the best approaches to teaching this aspect of science.

ACKNOWLEDGMENTS

We thank Mark Salata for initial suggestions about analyzing student questions, Douglas R. Brewster for support conducting this investigation, and David Keeling and three anonymous reviewers for critical review of this manuscript and constructive suggestions for improvement.

REFERENCES

Arzi, H. J., and White, R. T. (1986). Questions on students' questions. Res. Sci. Educ. *16*, 82–91.

Barden, L. M. (1995). Effective questioning and the ever-elusive higher-order question. Am. Biol. Teach. *57*, 423–426.

Bland, J. M., and Altman, D. G. (1995). Multiple significance tests: the Bonferroni method. Br. Med. J. *310*, 170.

Brill, G., and Yarden, A. (2003). Learning biology through research papers: a stimulus for question-asking by high school students. Cell Biol. Educ. *2*, 266–274.

Costa, J., Caldeira, H., Gallastegui, J. R., and Otero, J. (2000). An analysis of question asking on scientific texts explaining natural phenomena. J. Res. Sci. Teach. *37*, 602–614.

D'Agostino, R. B., and Stephens, M. A. (1986). Goodness-of-Fit Techniques, New York: Marcel Dekker.

Dillon, J. T. (1988). The remedial status of student questioning. J. Curric. Stud. *20*, 197–210.

Dori, Y. J., and Herscovitz, O. (1999). Question-posing capability as an alternative evaluation method: analysis of an environmental case study. J. Res. Sci. Teach. *36*, 411–430.

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

Ingram, E. L., Lehman, E., Love, A. C. and Polacek, K. M. (2004). Fostering inquiry in nonlaboratory settings: creating student-centered activities. J. Coll. Sci. Teach. *34*, 39–43.

Marbach-Ad, G., and Claassen, L. A. (2001). Improving students' questions in inquiry labs. Am. Biol. Teach. *63*, 410–419.

Marbach-Ad, G., and Sokolove, P. G. (2000a). Good science begins with good questions: answering the need for high-level questions in science. J. Coll. Sci. Teach. *30*, 192–195.

Marbach-Ad, G., and Sokolove, P. G. (2000b). Can undergraduate biology students learn to ask higher level questions? J. Res. Sci. Teach. *37*, 854–870.

National Research Council (1996). National Science Education Standards, Washington, DC: National Academies Press.

Pearson, J. C., and West, R. L. (1991). An initial investigation of the effects of gender on student questions in the classroom: developing a descriptive base. Commun. Educ. *40*, 22–32.

Polacek, K. M., and Keeling, E. L. (2005). Easy ways to promote inquiry in a laboratory course. J. Coll. Sci. Teach. *35*, 52–55.

Rosenshine, B., Meister, C., and Chapman, S. (1996). Teaching students to generate questions: a review of intervention studies. Rev. Educ. Res. *66*, 181–221.

Shodell, M. (1995). The question-driven classroom: student questions as course curriculum in biology. Am. Biol. Teach. *57*, 278–281.

Siegfried, T. (2005). In praise of hard questions. Science *309*, 76–77.

Watts, M., Gould, G., and Alsop, S. (1997). Questions of understanding: categorising pupils' questions in science. Sch. Sci. Rev. *79*, 57–63.

West, R., and Pearson, J. C. (1994). Antecedent and consequent conditions of student questioning: an analysis of classroom discourse across the university. Commun. Educ. *43*, 299–311.

Zar, J. H. (1999). Biostatistical Analysis, 4th ed., Upper Saddle River, NJ: Prentice Hall.

Zoller, U. (1987). The fostering of question-asking capability: a meaningful aspect of problem-solving in chemistry. J. Chem. Educ. *64*, 510–512.