Title

Critical Analysis of Primary Literature in a Master’s-Level Class: Effects on Self-Efficacy and Science-Process Skills

Permalink
https://escholarship.org/uc/item/7sd720gf

Journal
CBE—Life Sciences Education, 14(3)

ISSN
1931-7913

Authors
Abdullah, Christopher
Parris, Julian
Lie, Richard
et al.

Publication Date
2015-09-01

DOI
10.1187/cbe.14-10-0180

Peer reviewed

CBE—Life Sciences Education
Vol. 14, 1–13, Fall 2015

Article

Critical Analysis of Primary Literature in a Master's-Level Class: Effects on Self-Efficacy and Science-Process Skills

Christopher Abdullah,[*] Julian Parris,[†] Richard Lie,[‡] Amy Guzdar,[§] and Ella Tour[‖]

[*]Biomedical Sciences Graduate Program, [†]Department of Psychology, [‡]Department of Neurosciences, and [§]Section of Neurobiology and [‖]Section of Cell and Developmental Biology, Division of Biological Sciences, University of California, San Diego, La Jolla, CA 92093

Submitted October 21, 2014; Revised May 31, 2015; Accepted June 9, 2015
Monitoring Editor: Nancy Pelaez

The ability to think analytically and creatively is crucial for success in the modern workforce, particularly for graduate students, who often aim to become physicians or researchers. Analysis of the primary literature provides an excellent opportunity to practice these skills. We describe a course that includes a structured analysis of four research papers from diverse fields of biology and group exercises in proposing experiments that would follow up on these papers. To facilitate a critical approach to primary literature, we included a paper with questionable data interpretation and two papers investigating the same biological question yet reaching opposite conclusions. We report a significant increase in students' self-efficacy in analyzing data from research papers, evaluating authors' conclusions, and designing experiments. Using our science-process skills test, we observe a statistically significant increase in students' ability to propose an experiment that matches the goal of investigation. We also detect gains in interpretation of controls and quantitative analysis of data. No statistically significant changes were observed in questions that tested the skills of interpretation, inference, and evaluation.

INTRODUCTION

Rapid technological and scientific advances of the past few decades have generated demands for a workforce that possesses the skills associated with critical and creative scientific thinking: analysis and evaluation of data, problem solving, and generation of new concepts and ideas (Autor *et al.*, 2003; Autor and Price, 2013). These skills are often referred to as science-process skills (Coil *et al.*, 2010). Calls for an increased emphasis on teaching science-process skills have been issued by several major educational bodies in the recent past: the National Research Council (NRC, 2009), the American Association of Medical Colleges and Howard Hughes Medical Institute (AAMC-HHMI, 2009), and the American Association for the Advancement of Science (AAAS, 2011). For example, the *Vision and Change in Undergraduate Education: A Call for Action* report proposes that the ability to apply the process of science, described as "posing problems, generating hypotheses, designing experiments, observing nature, testing hypotheses, interpreting and evaluating data, and determining how to follow up on the findings," constitutes the first of the six fundamental core competencies that need to be developed by all undergraduate students (AAAS, 2011, p. 14).

Several studies have shown that despite overwhelming agreement that critical-thinking and science-process skills are very important instructional goals, very few college faculty members explicitly teach and assess these skills (Paul *et al.*, 1997; Coil *et al.*, 2010). Among the identified barriers to teaching these skills in biology classrooms are time constraints, the need to cover content, and the lack of validated, biology-specific assessments of critical thinking (Bissell and Lemons, 2006; Coil *et al.*, 2010). However, successful approaches to teaching critical-thinking and

CBE Life Sci Educ September 2, 2015 14:ar34
DOI:10.1187/cbe.14-10-0180
Address correspondence to: Ella Tour (etour@ucsd.edu).

science-process skills have been reported (e.g., Kitchen *et al.*, 2003; Dirks and Cunningham, 2006; Hoskins *et al.*, 2007; Coil *et al.*, 2010; Gottesman and Hoskins, 2013). The common theme in these studies is the implementation of a variety of active-learning approaches that include frequent practice of science-process skills inside and outside the classroom. However, such educational approaches are not common. Arguably, college-level biology education remains centered primarily around instructor-mediated transfer of facts (Alberts, 2009).

The need for critical-thinking and science-process instruction is even more acute in graduate education. Individuals with graduate degrees in biology tend to seek jobs that require routine use of higher-order thinking skills (e.g., physicians, researchers in academia and industry, educators). The need for physicians well-trained in problem solving, evaluating "competing claims in the medical literature and by those in medical industries" and capable of "application of scientific knowledge and scientific reasoning based on evidence" was articulated in the report *Scientific Foundations for Future Physicians* (AAMC-HHMI, 2009, pp. 4–5). Furthermore, individuals equipped with such skills face better job prospects. While the contribution of routine manual and routine cognitive skills (cognitive skills that can be replaced by computers, such as bookkeeping, clerical work) to the U.S. labor market has declined in the past 50 yr, the contribution of the nonroutine cognitive tasks (those that require critical-thinking and science-process skills) has been on the rise (Autor *et al.*, 2003; Autor and Price, 2013).

To assess students' science-process skills, we first need to define the different components of this complex set of skills. Bloom's taxonomy of educational objectives (Bloom *et al.*, 1956) provides a framework frequently used by educators for identifying the different components of science-process skills, designing activities, and creating assessments to evaluate these skills (Bissell and Lemons, 2006; Crowe *et al.*, 2008). Bloom's taxonomy identifies six categories of learning: knowledge, comprehension, application, analysis, synthesis, and evaluation (Bloom *et al.*, 1956). The first two categories are aligned with lower-order cognitive skills (LOCS), while the last three categories require higher-order cognitive skills (HOCS)—skills overlapping with critical thinking and science process (Zoller *et al.*, 1995; Crowe *et al.*, 2008; Coil *et al.*, 2010). The third category, application, is considered to be transitional between LOCS and HOCS. Another perspective for classifying the critical-thinking component of science-process skills is provided by the Delphi report of the American Philosophical Association (Facione, 1990). This report describes the consensus core critical-thinking skills, as determined by experts in the critical-thinking field, primarily from philosophy, social sciences, and education. According to the Delphi report, the core critical-thinking skills consist of interpretation, analysis, evaluation, inference, explanation, and self-regulation (Facione, 1990). Importantly, the Delphi report also notes the overlap between the different categories of critical thinking, suggesting that "creating arbitrary differentiation simply to force each and every subskill to become conceptually discrete from all others is neither necessary nor useful" (Facione, 1990, p. 6). For example, to evaluate a hypothesis, one needs to analyze the data on which the hypothesis is based and draw one's own conclusions from these data. The frameworks of Bloom's taxonomy and the Delphi report provide useful complementary perspectives for classification of science-process skills. For example, such an important science-process skill as experimental design is not included in the core consensus critical-thinking skills defined by the Delphi report (Facione, 1990), while in Bloom's taxonomy it is categorized as synthesis, one of the HOCS (Bloom *et al.*, 1956; Crowe *et al.*, 2008).

In classroom settings, discussion of primary literature provides an excellent opportunity to practice science-process skills: analyzing the data presented, drawing independent conclusions, evaluating the authors' conclusions, synthesizing new hypotheses, and designing new experiments to test them. Several studies have reported that undergraduate courses that focus on analysis of primary literature have positive effects on students' science-process and critical-thinking skills (Hoskins *et al.*, 2007; Gottesman and Hoskins, 2013; Segura-Totten and Dalman, 2013). For example, the CREATE (Consider, Read, Elucidate hypothesis, Analyze and interpret the data, and Think of the next Experiment) approach, which offers structured engagement with linked sequences of articles from the same lab, was associated with a statistically significant increase in students' skills in data analysis and drawing logical conclusions in an upper-division undergraduate class (Hoskins *et al.*, 2007) and significantly improved students' performance in the Critical Thinking Assessment Test (CAT; Stein *et al.*, 2012) in a freshmen-level class (Gottesman and Hoskins, 2013). However, the effects of other primary literature–centered approaches on the development of science-process skills, in particular among graduate students, remain unexplored.

We report here on the development and assessment of a primary literature–based course designed for students enrolled in the contiguous BS/MS program in biology at the University of California, San Diego (UCSD). In this research-based master's program, biology undergraduates can extend the research they perform in their senior undergraduate year to obtain a master's degree. As graduate students, they are routinely expected to use various science-process skills: interpreting primary literature, contributing to experimental design, analyzing results, and, finally, writing and defending a substantial research thesis within 1 or 2 yr after graduating with a bachelor's degree. However, as we will demonstrate here, our master's students often feel unprepared for these tasks. One of our goals was to design a course that can improve their skills of critical analysis of primary literature and experimental design and increase their sense of self-efficacy in their ability to perform these tasks.

To achieve these goals, the course described here incorporated structured group and individual activities in which students practiced skills required to understand and analyze four papers from diverse fields of biology. Students' evaluations point to significant perceived gains in science-process skills in the context of primary literature. However, using a science-process skills test, we detected a statistically significant increase in students' ability to propose an experiment that matches the goals of investigation, interpretation of controls, and quantitative analysis of data, but we did not see a similar increase in responses to questions assessing the skills of inference and evaluation.

METHODS

Students' Demographics and Career Aspirations

The elective course described in this study was designed and offered specifically to the students enrolled in the contiguous BS/MS program of the Division of Biological Sciences at the UCSD. Only UCSD biology undergraduates can enter this program during their senior year. Altogether, they complete at least six quarters of research (typically, three quarters as undergraduates, followed by at least three quarters of graduate research in the same lab) and defend a research-based thesis.

The data on students' demographics, major, career aspirations, and experience with primary literature were collected via anonymous surveys. Figure 1 presents the data collected in Fall 2013 and Winter 2014 from 28 students who completed the beginning-of-the-quarter survey. The majority of students participating in this study were recent undergraduates, primarily in their first (59%) or second (25%) quarter of the master's program (Figure 1A). Forty-five percent of our students were Asian, another 45% were non-Hispanic white, and 3% were Hispanic or Latino (Figure 1B). The students represented all UCSD biology undergraduate majors, except for bioinformatics (Figure 1C). Medicine was the most frequently considered career aspiration, followed by biotechnology and teaching (Figure 1D).

Selection of Scientific Papers

The course described in this study, BGGN 211: Recent Advances and Experimental Approaches in Modern Biology, was taught by one of the authors (E.T.). Because this is the only course that is specifically geared toward master's students, it was designed to have an appeal to master's students working in a variety of subfields of biology. With this goal in mind, we selected the course papers to increase the variety of experimental approaches and to train students in critical analysis of papers outside their areas of expertise (Supplemental Table S1). The papers were also chosen to achieve our educational goal (Muench, 2000), namely, to enhance the skills of critical analysis of scientific literature (Figure 2A). The first paper used relatively straightforward experimental techniques and, importantly, contained drawbacks in experimental design and occasional flaws in data interpretation that did not require expert knowledge to detect. The second paper was selected by the students (via online voting) out of a group of articles from different fields in biology that were suggested by local biology faculty members as well-designed and important recent publications in their areas of research. The third and fourth papers were research articles that addressed the same experimental question but reached opposite conclusions (Figure 2A). These conflicting papers were included to prompt students to critically

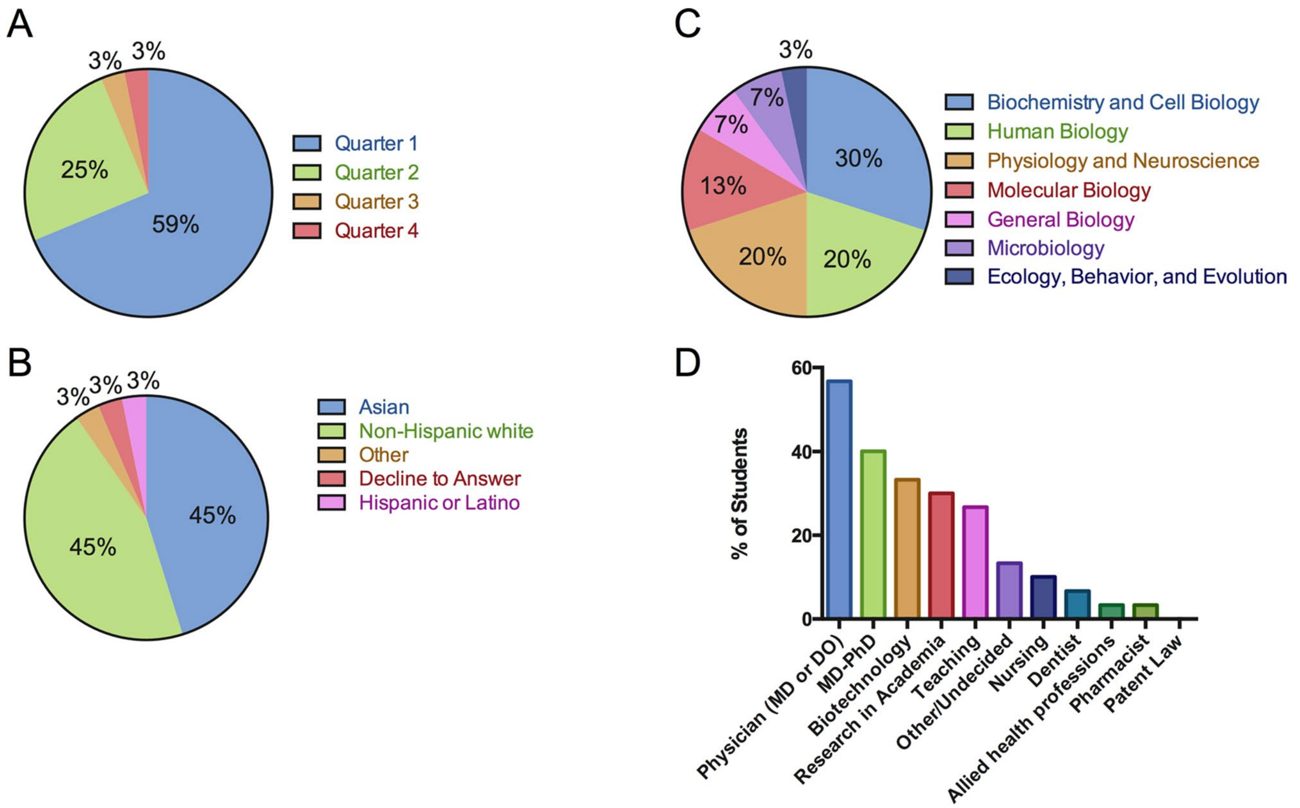

Figure 1. Students' demographics and career aspirations, based on an anonymous precourse survey. $n = 28$ students. (A) Quarter in the master's program. (B) Students' ethnic background. (C) UCSD biology major affiliation as undergraduate. (D) Students' current career aspirations. Students could select all career options they are currently considering from a list of options, so the sum of responses exceeds the total number of students.

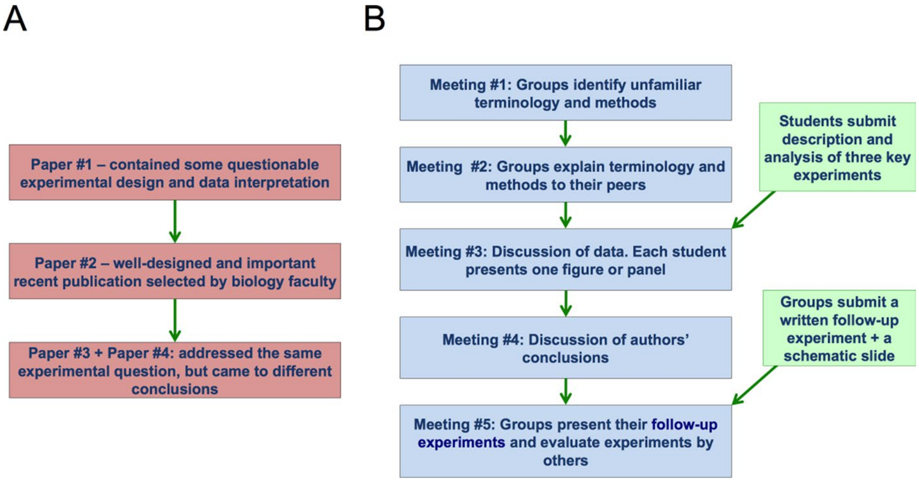
A
Paper #1 – contained some questionable experimental design and data interpretation
Paper #2 – well-designed and important recent publication selected by biology faculty
Paper #3 + Paper #4: addressed the same experimental question, but came to different conclusions
B
Meeting #1: Groups identify unfamiliar terminology and methods
Meeting #2: Groups explain terminology and methods to their peers
Meeting #3: Discussion of data. Each student presents one figure or panel
Meeting #4: Discussion of authors' conclusions
Meeting #5: Groups present their follow-up experiments and evaluate experiments by others
Students submit description and analysis of three key experiments
Groups submit a written follow-up experiment + a schematic slide

Figure 2. Course modules and individual module structure. (A) Papers discussed in the three course modules. The first two modules focused on one paper, while the third module focused on two papers that addressed the same experimental question but reached different conclusions. (B) The structure of a module. Each module consisted of five class meetings. Assignments outside class included: written analysis of three key experiments (submitted individually) and a follow-up experimental proposal that was submitted by groups of three to five students.

examine every aspect of the two papers to try to determine which group of authors had a stronger scientific argument.

Course Activities

The course met for 80 min twice a week over a 10-wk quarter. It had three modules, each module centered on the focus paper(s), with five meetings in each module (Figure 2B). We designed the modules to provide a stepwise, structured approach to a paper. Anonymous surveys from previous quarters indicated that unfamiliar background and unfamiliar experimental techniques were among the most difficult aspects of scientific papers for our students. To provide students with practice in overcoming these difficulties, during the first meeting of each module, we had student work in small groups to identify the important methods, terminology, and background they would need to know to understand the paper. Each group was assigned one or two of the identified techniques and background items, which they presented to the class in the second meeting of the module (Figure 2B).

Before the third meeting, students read the entire assigned paper and wrote individual analyses of three key experiments (as selected by the student). Guidelines for this assignment prompted the students to provide a detailed description of experimental setups and their own analyses of these experiments (Appendix A in the Supplemental Material). During the third meeting of the module, most students were asked to present at least one experiment from the paper, focusing on clearly describing how the experiment was performed and on their own interpretation of the data. Each figure presentation was followed by a brief question-and-answer session. The entire class was encouraged to participate in both asking and answering questions. The instructor also asked probing questions, prompting students to question authors' interpretations and conclusions, evaluate authors' experimental design, and decide whether additional controls were needed. Asking a thoughtful question and, especially, providing an answer to a question earned the students participation points. Each student had to accumulate at least 20 participation points to earn a perfect score in the "participation" category (accounting for 20% of the overall grade). These participation points seemed to provide students with an additional motivation to participate in the discussion, especially in the beginning of the course. In the second half of the course, participation in discussions became habitual for most of the students.

The fourth class meeting was dedicated to the discussion of the authors' conclusions and identification of questions that remain to be answered. The groups also worked on designing an experiment that would follow up on the paper. Before the last meeting of the module, each group of students submitted a written one-page proposal of the follow-up experiment. The guidelines for this assignment are provided in Appendix B in the Supplemental Material. Briefly, the assignment included articulating the experimental question and its importance, providing a detailed experimental design, including the controls, and then predicting the expected outcomes. Each group also provided a slide that contained a schematic of its experiment. During the last class meeting of each module, the groups presented their experimental proposals and evaluated proposals presented by other groups, acting as "grant panels," an activity described in the CREATE approach (Hoskins *et al.*, 2007; Hoskins and Stevens, 2009). Each experimental design presentation was followed by a brief question-and-answer session.

Science-Process Test

To assess the progress of our students in science-process skills, we developed a test in which students were asked to interpret data from two experiments, evaluate hypotheses based on these experiments, and propose their own experimental designs (Appendices C and D in the Supplemental Material). Because scientists use more than one approach to investigate a question, our test included two related experiments, one being a follow-up to the other. The experimental approaches and data presentation in this test were selected to minimize any specialized background knowledge required to understand the questions. The experimental condition utilized RNA interference–mediated knockdown of a target gene. While this technique may not be familiar to all the students, descriptions of the technique and its effects on expression of a gene were provided in the prompt. The students were asked to consider two pieces of data that reflected the effects of a down-regulation of a hypothetical gene (gene X or Y) on cell numbers and programmed cell death (apoptosis). In the first part of the test (questions 1-1 through 1-3, Appendices C and D in the Supplemental Material), the students were asked to interpret data, draw conclusions, and evaluate a hypothesis based on the first piece of data. In the second part of the test (questions 2-1 through 4-1), the second piece of data was presented. Questions 2-3, 3-1, and 4-1 then prompted students to evaluate a hypothesis (question 2-3) and propose a new hypothesis based on both pieces of data (questions 3-1 and 4-1). Finally, in the third part of the test (question 4-2), the students were asked to design an experiment to test a hypothesis that a particular mutation in gene X (or gene Y) contributes to cancer development. Importantly, at this point, students could use the experimental approaches they saw in part 1 and 2 of the test as the basis for their own experimental design, thus minimizing the need for specialized knowledge in how cells can be manipulated. On the other hand, the students were also free to choose a completely different experimental approach.

Two isomorphic versions of the test were generated (Appendices C and D in the Supplemental Material). Development of the test was an iterative process, wherein the authors used comments from biology PhD students who took the test and data from the two quarters in which the test was piloted (Fall 2012 and Winter 2013) to revise and clarify the questions. The version used here (Fall 2013 and Winter 2014) was reviewed by three experts in biology education (who also had PhDs in biology) from three different institutions and was revised based on their comments. Finally, the alignment of the individual questions in our science-process skills test with the consensus critical-thinking skills (Facione, 1990) was conducted by nine biology faculty members at three different institutions and three postdocs and three graduate students who are members of the UCSD STEM-Education and Diversity Discussion group. The validators were provided with brief descriptions of each of the consensus critical-thinking skills as defined in the Delphi report and with the text that contained the full consensus descriptions of core critical-thinking skills and subskills (Facione, 1990). The validators were asked to select all core critical-thinking skills that were required to answer each of question in our test. The results of this survey for all questions, except for the experimental design question (Q4-2), are summarized in Supplemental Table S2. For each question, the core critical-thinking skill that received the most votes was designated as the primary skill, while the skill that received 50% or more votes was designated as the secondary skill.

Although not considered to be one of the consensus core critical-thinking skills in the Delphi report (Facione, 1990), experimental design is one of the core science-process skills and it aligns with the Bloom's category of synthesis. The experimental design score was based on seven components that included appropriateness (the match between the hypothesis and the proposed experiment), identification of experimental system, treatment, control group, assay, quantity measured, and expected outcomes (Supplemental Table S3). To decrease the probability of students unintentionally omitting aspects of experimental design, we included these elements in the prompt of the experimental design question. During the writing of this article, a paper by Dasgupta and colleagues was published that described a comprehensive review of the difficulties in experimental design that had been described in the K–12 and college education literature and that also provided a rubric of experimental design (RED) that targeted these identified difficulties (Dasgupta *et al.*, 2014). In Supplemental Table S3, we match the experimental design categories scored in our test with the experimental design difficulties identified by Dasgupta and colleagues and the difficulties assessed in the RED (Dasgupta *et al.*, 2014).

Science-Process Test Administration and Scoring

The tests were administered in class during week 1 (pretest) and week 10 (posttest) in a counterbalanced design, such that half of the students were randomly assigned to take version A as a pretest and version B as the posttest, and vice versa. Thirty-three students took both pre- and posttests in Fall 2013 and Winter 2014 quarters. The tests were deidentified, and each test received a randomly generated number. The tests were then evaluated by three of the authors, who are biology faculty members or graduate students in biology (C.A., E.T., and R.L.). The raters were blind to both students' identities and to the pre/post status of the test. A random sample of approximately 10 tests of each version was selected as a training set, and the scoring rubric was developed based on this sample. The three raters scored the entire training set together, discussing each score (Appendix E in the Supplemental Material). Each rater then scored the remaining tests independently. In cases in which the score for a particular question differed by 50% or more among the raters, all three raters re-examined the student's response and discussed the ratings. The goal of these face-to-face discussions was to make sure that the student's handwritten response was correctly read and a consensus interpretation was achieved. Each of the raters articulated his or her reasoning for giving the response a particular rating. In some cases, but not in all, these discussions lead the raters to revise their scores. Interrater reliability was high after the revision process (Cronbach alpha > 0.90 across 27 items). The tests were then reidentified, and pre- and posttests were matched. Dependent-measures *t* tests were used to assess changes in student performance between the pre- and posttest.

Students' Anonymous Surveys

Students' anonymous surveys were administered online via SurveyMonkey (www.surveymonkey.com), during the first and the last week of a 10-wk quarter. Both pre- and postversions of the surveys assessed students self-efficacy in skills related to critical analysis of primary literature (for the full list of self-efficacy questions, see Supplemental Table S2). The preinstruction survey also contained questions about students' demographics and career aspirations (Figure 1). The students received small course credit for completing the surveys. To allow matching between the beginning- and the end-of-the-quarter surveys while preserving the anonymity of responses, we asked students to provide the same five-digit number in both surveys. Twenty-eight students completed both pre- and postsurveys in the Fall 2013 and Winter 2014 quarters. Wilcoxon signed-rank tests were used to analyze the changes in students' self-efficacy ratings.

Institutional Review Board

Protocols used in this study were approved by UCSD Human Research Protections Program (project 111351SX).

RESULTS

Students' Self-Efficacy

To gain insight into whether students perceived any change in their science-process skills as a result of taking this course, we administered an anonymous online survey at the beginning and the end of the quarter. Using a five-point Likert scale ranging from "poor" to "excellent," we asked the students to rate their current skills in interpretation and inference (interpreting data from a paper and independently drawing conclusions), evaluation (critically evaluating authors' conclusions), and experimental design (proposing an experiment with the appropriate controls as follow-up on a paper). Thus, our survey provided us with a readout of students' self-efficacy: "the construct of perceived confidence in executing a given behavior" (Baldwin *et al.*, 1999). Because students' self-efficacy could depend on whether or not the paper was from the area of their master's research, we asked the students to evaluate their skills for a paper within and outside their research areas separately (Table 1). In Figure 3, we present the combined results (skills within and outside students' areas of expertise) of 28 pairs of students' responses, grouped into categories of interpretation and inference, evaluation, and experimental design. Supplemental Table S4 contains students' self-efficacy ratings for all individual questions.

At the beginning of the quarter, only 58% of the students rated their skills in interpretation and inference as good (32%), very good (19%), or excellent (7%; Figure 3). A substantial positive shift was observed at the end of the quarter, with 87% of the students rating their interpretation and inference skills as good (29%), very good (37%), or excellent (21%; Figure 3). The difference in students' self-efficacy ratings in analysis was statistically significant (Wilcoxon signed-rank test, $S = 89.50$, $p < 0.0001$, Cohen's $d = 1.11$, $n = 28$). In the categories of evaluation and experimental design, the students' ratings in the presurvey were lower than in the category of interpretation and inference. In the category of evaluation, 53% of the students rated their skills low in the beginning of the quarter, with 20% of the students rating their skills as poor and 33% as adequate (Figure 3). A statistically significant increase in students' self-efficacy in this category was observed at the end of the quarter ($S = 105.00$, $p < 0.0001$, Cohen's $d = 1.27$), at which time none of the students rated their skills as poor, and only 18% rated them as adequate (Figure 3). Similar trends were observed in students' self-efficacy ratings in experimental design: at the beginning of the quarter, 59% of the students rated their skills either as poor (29%) or adequate (30%; Figure 3). A significant shift ($S = 120.00$, $p < 0.0001$, Cohen's $d = 1.19$) in students' self-efficacy occurred at the end of the quarter: only 20% rated their skills as poor (2%) or adequate (18%; Figure 3).

Students rated their skills higher when asked about a paper within their areas of research, as opposed to a paper outside their areas of research (Supplemental Figure S1). For example, in the beginning of the quarter, when asked about their self-efficacy in proposing an experiment following up on a paper *outside* their areas of research, none of the students evaluated their skills as very good or excellent (Supplemental Figure S1A). At the same time, 28% of students evaluated their skills of proposing an experiment that would follow up on a paper *within* their areas of research as either very good (21%) or excellent (7%; Supplemental Figure S1A). At the end of the quarter, a statistically significant increase in

Table 1. List of questions used to assess students' self-efficacy in science-process skills in the context of scientific papers[a]

Interpretation and inference
Interpreting data in a paper *within* your area of research
Interpreting data in a paper *outside* your area of research
Independently drawing conclusions from data presented in a paper *in* your area of research
Independently drawing conclusions from data presented in a paper *outside* your area of research
Evaluation
Critically evaluating authors' conclusions in a paper *in* your area of research
Critically evaluating authors' conclusions in a paper *outside* your area of research
Experimental design
Proposing an experiment, with the appropriate controls, that would follow up on a paper *in* your area of research
Proposing an experiment, with the appropriate controls, that would follow up on a paper *outside* your area of research

[a] Questions were grouped in the categories of interpretation and inference, evaluation, or experimental design, as indicated.

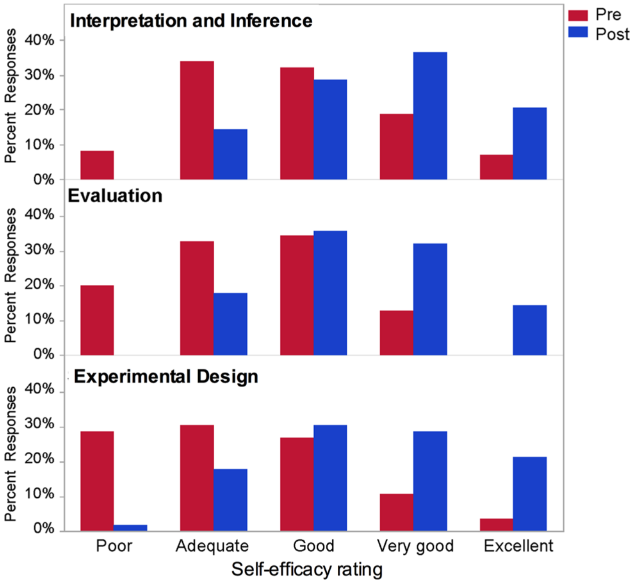

Figure 3. Students' self-efficacy in science-process skills in the context of primary literature. Twenty-eight pairs of matched responses from anonymous surveys given at the beginning (Pre) and end (Post) of the quarter were analyzed. A list of survey questions that were grouped into the categories of interpretation and inference, evaluation, and experimental design is provided in Table 1. "Percent responses" refers to the frequency of a specific rating (poor, adequate, etc.) among all responses to the questions that were grouped into the same category (interpretation and inference, evaluation, or experimental design).

self-efficacy in proposing follow-up experiments to papers both within and outside of students' areas of research was observed (within: $S = 79.00$, $p < 0.0001$; outside: $S = 99.50$, $p < 0.0001$). After instruction, 36% of students evaluated their self-efficacy in proposing an experiment that would follow up on a paper *outside* their areas of research either as very good (29%) or excellent (7%; Supplemental Figure S1A). When asked about proposing experiments *in their own* fields, 64% rated their skills at the synthesis level as either very good (29%) or excellent (36%; Supplemental Figure S1A).

Analysis of Science-Process Skills Test

Students' surveys pointed to high gains in their *perceived* level of science-process skills in the context of analysis of scientific papers; however, we wished to examine whether we could also detect measurable changes in student performance in science-process skills in the context of biological experiments. To that end, we designed two isomorphic versions of a test in which students were presented with a sequence of two experiments that examined the same experimental system (Supplemental Material, Appendices C and D). In the test, students were asked to analyze data, draw conclusions, and evaluate and propose hypotheses based first on one piece of data and then on both pieces of data, and to design a follow-up experiment (see Supplemental Table S2 for alignment of the questions with the consensus core critical-thinking skills and Supplemental Table S3 for the components of the experimental design score). The data in the test were presented and described in such a way as to minimize the demands for a specialized subject and technical knowledge in any specific area of biology. Thirty-three paired (pre- and postinstruction) tests were rated by three raters blind to both the identity of the students and the pre/post status of the test (the scoring rubric is provided in Appendix E in the Supplemental Material). Dependent-measures *t* tests were used to assess whether preinstruction and postinstruction scores in each of the categories were statistically different. No statistically significant changes in posttests were observed in the categories of interpretation, inference, and evaluation (Figure 4A). Statistically significant gains were detected in the experimental design category ($p = 0.039$, Figure 4A). More detailed analysis of student performance in each category is presented below.

In the interpretation category, students scored very high in both the pre- and posttest (89.9% and 94.1%, respectively), indicating the data presented in the test were accessible to the vast majority of the students (Figure 4A). The inference category questions probed for two types of skills: drawing conclusions from two pieces of experimental data presented in the test and proposing hypotheses (Appendix E in the Supplemental Material). The average pre- and postscores in the "drawing conclusions" subcategory were higher than the corresponding scores in the "proposing hypothesis" subcategory, implying that students found the latter skill more challenging (Supplemental Figure S2A). Positive but not statistically significant trends were observed in student performance postcourse in both subcategories (Supplemental Figure S2A).

Questions aligned with the evaluation skill asked students to evaluate a hypothesis, based first on one piece of data and then on two pieces of data. Not surprisingly, the latter task was more challenging to the students: in the pretest, the average score for the evaluation question based on one piece of data was 81%, while the average score for the evaluation question based on two pieces of data was 66.3%, a 14.7% difference (Supplemental Figure S2B). At the end of the course, the difference decreased to only 2.4% (Supplemental Figure S2B); however, this change was not statistically significant ($p = 0.191$).

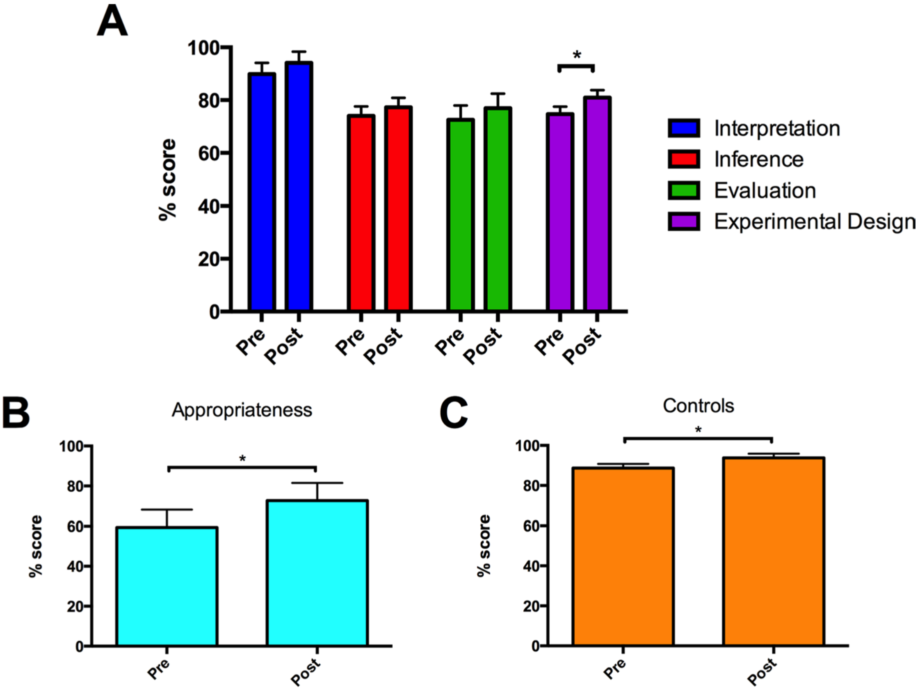

Figure 4. Analysis of science-process tests. Thirty-three pairs of matched tests from the beginning (Pre) and end (Post) of the quarter were evaluated. (A) Student performance in the categories of interpretation, inference, evaluation, and experimental design, before and after instruction. Small but statistically significant gains were observed in the experimental design category ($p = 0.039$, Cohen's $d = 0.379$). (B) Average pre- and posttest score in the category of appropriateness. The increase in the posttest scores was statistically significant ($p = 0.005$, Cohen's $d = 0.526$). (C) Average pre- and posttest score in the category of interpretation of controls. The increase in the posttest scores was statistically significant ($p = 0.049$, Cohen's $d = 0.37$).

A statistically significant increase was observed in students' ability to propose an experiment based on a given hypothesis (p = 0.039, Cohen's d = 0.379). A similar statistically significant increase in experimental design ability was observed in the two quarters in which this test was piloted (Fall 2012 and Winter 2013): 61.8% for the mean pretest score and 72.2% for the mean posttest score (p = 0.011, n = 42; Supplemental Figure S4A). In the 2012–2013 version of the experimental design question, the follow-up experiment was not constrained to a specific hypothesis: students were asked to propose any experiment that would follow up on the experiments presented in the test.

Experimental Design Ability

In the experimental design part of our science-process test, students were given a scenario that described a large family with high incidence of certain cancer. Genomic sequencing of the family members revealed a correlation between the presence of a particular mutation in the gene interrogated in the first part of the test and the development of brain cancer (version A) or colon cancer (version B). Students were asked to design an experiment that would test the hypothesis that this mutation contributes to cancer development. As in a real scientific investigation, more than one appropriate experiment could be proposed. The experimental design score consisted of seven components: appropriateness to the goal of investigation, identification of an experimental system, treatment, assay, quantity measured, identification of controls, and statement of anticipated outcomes (see Supplemental Table S3 for the alignment of these components with the previously described difficulties in experimental design identified in Dasgupta *et al.*, 2014). The experimental design component in which the students had the lowest pretest mean score (59.5%) was appropriateness of the proposed experiment to the goal of investigation (Figure 4B and Supplemental Figure S3). Examples of quotes taken from students' responses relevant to the appropriateness of the experimental design are shown in Table 2. Experiments that appropriately addressed the given hypothesis (a specific mutation in gene X or Y contributes to brain or colon cancer development) ranged from introducing this mutation into animals and looking for tumor development (example 1272) to transfecting the gene with the mutation into brain or colon cells and looking for increased cell proliferation (example 8529). Partially correct or incorrect responses proposed

Table 2. Example quotes of students' responses relevant to the "appropriateness" of the proposed experimental design to the goal of investigation[a]

Test number	Student response	Quality of response/ rater's comments	Score (out of 2)
1272	"In vivo assay testing the proliferation effects of Gene Y mutation on tumor formation. Gene Y mutation containing colon cancer cells will be injected subcutaneously *in different amounts* into immunodeficient mice and monitored biweekly for tumor formation. Controls: inject healthy colon cells without gene Y mutation into mice at some place."	Appropriate experiment. Note that the italicized text indicates a problem with combinatorial reasoning (Dasgupta *et al.*, 2014): different numbers of cells are injected in the treatment but not in the control condition. This aspect of the experimental design is evaluated in the "treatment/independent variable" category.	2
8529	"The assay would be whether or not the rats develop colon cancer … We would be observing the appearance of colon cancer in rats that are just past middle aged. Controls would include: healthy rats with no gene Y mutation; gene Y mutation rats."	Appropriate experiment. Note that the student considers the treatment condition to be one of the controls.	1.67
6582	"Use cultured cells to introduce the same Gene Y mutation as seen in humans … perform a proliferation assay to measure the number of proliferated cells."	Partially correct response. The response describes an appropriate experiment; however, it does not address the colon cancer aspect of the prompt.	1.5
3064	"I would perform an experiment on cultured cells to overexpress Gene X and look at its effects by transfecting cells with an expression vector that contains Gene X and a strong promoter … A proliferation assay would be performed on all the controls and the Gene X overexpression cells."	Inappropriate experiment for the given hypothesis (does *the specific mutation* in gene X contribute to cancer). The experiment proposed by the student addresses a related question: Does overexpression of gene X contribute to cancer?	1
1524	"Transfection of Gene Y into non-Gene Y expressing cells to look for increased proliferation."	Inappropriate experiment for the given hypothesis. The experiment proposed here tests a related question: Does the expression of Gene Y contribute to proliferation?	0.5
7803	"Clinical study w/human patients to see if there are any individuals w/o gene X mutant, but w/cancer."	Inappropriate experiment for the given hypothesis. The experiment proposed here tests a different question: Can cancer develop without a mutation in gene X?	0

[a]The scores are the average scores of three raters. The complete student responses quoted here are shown in Supplemental Table S5.

experiments that aimed to answer different, although related, questions, such as:

Example 3064: Does overexpression of gene X increase cell proliferation? (note that the effect of the specific mutation is not tested here)

Example 1524: Does expression of gene Y increase cell proliferation?

Example 7803: Can a person have cancer without having a mutation in gene X?

The subcategory of appropriateness also showed the largest and the only statistically significant increase in the posttest mean score (to 72.5%, $p = 0.005$, Cohen's $d = 0.526$; Figure 4B). Components of experimental design in which students performed very well in the pretest and did not show statistically significant change in the posttest were clear identification of experimental system ("experimental subject" in Dasgupta *et al.*, 2014) and inclusion of an appropriate control group (Supplemental Figure S3). No significant changes were observed in such components of experimental design as "independent variable/treatment," "quantity measured," and "expected outcomes" (Supplemental Figure S3). This lack of change could be due, in part, to the fact that the experimental design question prompted the students to include the experimental system, the assay (treatment), what will be measured, and the controls. A very similar science-process test used in Fall 2012–Winter 2013 quarters in this course did not specify which components of experimental design should be included, instead prompting the students to "include all relevant components of experimental design in your experiment." Importantly, the students were free to propose any experiment that would follow up on the data given in the test. Forty-one pairs of pre- and postquarter tests were scored by three raters who were blind to students' identities and the pre- or postcourse status of the test. Statistically significant gains were observed in "experimental system," "independent variable/treatment," "assay," and "quantity measured" components of experimental design (Supplemental Figure S5). The precourse scores in most components of experimental design were also substantially lower in this version of the test in comparison with the later version (compare Supplemental Figures S3 and S5).

Gains in Interpretation of Controls and Quantitative Analysis of the Data

In the analysis of the test questions aligned with interpretation, questions pertaining to identification and interpretation of controls were analyzed separately. Students' tests were scored based on their ability to correctly identify controls in the two experiments described in the test and explain why these controls were included (Figure 4C). The students performed very well in this category in the pretest: the average score in this category was 88.6%. A small but statistically significant increase was observed in this category in the posttest: the mean score was 93.8% ($p = 0.0491$, Cohen's $d = 0.37$; Figure 4C).

Analysis of the science-process tests also revealed statistically significant increases in students' average scores for quantitative analysis of the data (Figure 5). The score in the "quantitative data analysis" category was determined based on the presence and correctness of the comparison between experimental and control data in quantitative terms (e.g., percentages, fold difference; see Appendix E in the Supplemental Material). Throughout the course, students were encouraged to use quantitative terms both in their written paper analyses, where quantifications of differences between experiment and controls were part of the grade, and in the in-class discussions. We observed more than a twofold increase in the average postinstruction score (mean pretest score: 26.7%, mean posttest score: 60.0%, $p = 0.0002$, Cohen's $d = 0.728$). A similar increase was observed in the two quarters in which this test was piloted (Fall 2012 and Winter 2013, $p = 0.015$, Cohen's $d = 0.556$, $n = 41$; Supplementary Figure S4B).

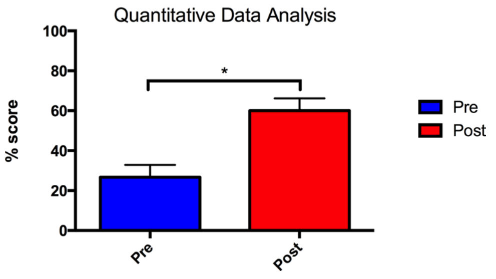

Figure 5. Average pre- and posttest scores in the quantitative data analysis category. The increase in the posttest scores was statistically significant ($p = 0.002$, Cohen's $d = 0.728$, $n = 33$ pairs of pre- and posttests).

DISCUSSION

The course described here utilized four primary research papers from diverse topics in biology, selected with the goal of providing students with opportunities to practice science-process skills. The first paper had drawbacks in experimental design and data interpretation. The second paper was an exemplary scientific investigation. The third and the fourth papers investigated the same experimental question but came to different conclusions. This instructional approach correlated with highly significant increases in students' self-efficacy in a variety of science-process skills: drawing conclusions from data presented in scientific papers, critically evaluating authors' conclusions, and designing an experiment that would follow up on a paper. Using a science-process test that we developed, we detected a statistically significant increase in students' ability to propose an experiment appropriate to the goal of investigation, describe data in quantitative terms, and interpret controls. However, we did not detect statistically significant increases in students' performance on questions aligned with data interpretation, inference, and evaluation.

Gains in Proposing an Experiment Appropriate to the Goal of Investigation

We detected a statistically significant increase in students' ability to propose an experiment that matched the given hypothesis. Analysis of students' responses indicated that students often came up with an experimental design that was testing a related but different hypothesis. Designing an experiment in which the treatment or the outcomes appropriately address a specific research question represent known areas of difficulty (Dasgupta *et al.*, 2014). We did not detect statistically significant gains in other components of experimental design, such as identification of experimental system, treatment, assay, quantity measured, identification of controls, and statement of anticipated outcomes. A possible contributor to this lack of observed increase is the fact that the experimental design prompt in the test directed the students to provide these components. The purpose of this inclusion was to uncover any ideas students might have about these components; however, this prevented us from assessing what components of experimental design students would provide without being prompted and may have limited our ability to detect changes in experimental design skills. Indeed, when an earlier version of this test that did not provide such prompts was used in this class (Fall 2012 and Winter 2013), statistically significant increases in "experimental system," "independent variable/treatment," and "quantity measured components" were detected. Our ability to compare between the two versions of the test is limited, however, by the fact that the earlier version prompted students to propose any experiment to follow up on the data described in the test, whereas in the later version, students were asked to design an experiment that would test a given hypothesis.

Several potential confounding factors could influence the gains we saw in the appropriateness of the proposed experiments. These include:

1. If students analyzed or designed similar experiments during this course, they might be expected to perform better in the posttest.
2. Students whose major was outside the cell/molecular biology field (e.g., ecology, behavior, and evolution majors who comprised 3% of our students) would be expected to gain knowledge in this field after taking this course and therefore perform better in the posttest.
3. Students' own lab research, as well as other courses they were taking during the same quarter, could positively contribute to their performance in the posttest.

While we acknowledge a possible and likely contribution of the second and the third confounders, we believe that the first confounder is unlikely to explain the gains we observed. None of the papers that we examined in this class involved cancer, which was the topic of the questions in our science-process test, and follow-up experiments proposed by the students during the course were therefore different from those proposed by students in the test. We sought to minimize the effects of the second confounder, subject knowledge gain, by selecting experimental approaches and data presentation used in the test to minimize any specialized background knowledge in biology needed to understand the experiments (the students were expected to know, however, that it is possible to isolate or generate a mutant version of a gene and introduce it into a cell). The effect of familiarity with specific techniques in cell and molecular biology was minimized by the fact that students could use the experimental techniques described in the first part of the test (transfection of cells, cell counts, or apoptosis assays) and adopt them to their experimental design (this strategy was used in only two responses). Finally, the specific component of experimental design that significantly increased after instruction was the match between the proposed experiment and the experimental hypothesis provided. Increase in this category cannot be easily explained by increase in subject knowledge or familiarity with a particular experimental technique.

We could not directly compare our study with other studies that quantified the effects of interventions aimed to enhance the skills of experimental design, because these studies were conducted in different levels of classes (introductory and nonmajor) and utilized different assessment methods (Sirum and Humburg, 2011; Gottesman and Hoskins, 2013; Brownell *et al.*, 2014). Sirum and Humburg (2011) and Gottesman and Hoskins (2013) used the Experimental Design Ability Test (EDAT), an open-response test that measures students' ability to design an experiment testing a claim about the effectiveness of an herbal supplement (Sirum and Humburg, 2011). A student's response in EDAT is scored based on parameters incorporating the fundamental elements of experimental design (in contrast to our test, the EDAT prompt does not include directions as to which experimental design elements should be included). The Expanded EDAT (E-EDAT) uses a similar prompt, with modifications that include asking students to provide justification for their responses (Brownell *et al.*, 2014).

While EDAT and E-EDAT assess general, not subject-specific, experimental design skills, our test probed students' ability to design an experiment using discipline-specific knowledge expected from a graduate with a BS in biology. As we will argue below, the ability to apply discipline-specific knowledge and its "methodological principles" (Facione, 1990, p. 5) is essential to critical thinking within the discipline (Facione, 1990; Bailin *et al.*, 1999; Willingham, 2008). Such discipline-specific experimental design skills as identifying and explaining the purpose of controls in the context of complex biological experiments were the focus of the investigation by Shi and colleagues, who demonstrated that an online tutorial and seven in-class exercises resulted in significant gains in these skills in a sophomore-level cell biology lab course (Shi *et al.*, 2011).

Our assessment did not include such important elements of EDAT or E-EDAT as understanding that experiments have to be repeated, evaluation of the sample size, and knowledge that one can never unambiguously prove a hypothesis (Sirum and Humburg, 2011; Brownell *et al.*, 2014). However, our test included an assessment of the appropriateness of the proposed experiment to the hypothesis being tested, identification of an appropriate control group, and statement of outcomes that would support the hypothesis. In future studies, it will be interesting to compare students' performance in EDAT or E-EDAT and our revised test (in which the prompts of experimental design components will be removed).

Recently, two new tools have become available to assess experimental design abilities at more advanced levels (Dasgupta *et al.*, 2014; University of British Columbia, 2014). The RED, which uses open-response answers, can be used to examine the salient features of both content-specific and content-independent experimental design questions (Dasgupta *et al.*, 2014). The Experimental Design (Third/Fourth Year Undergraduate Level) Concept Inventory is a validated tool that allows examination of students' knowledge of experimental design in a multiple-choice format (University of British Columbia, 2014).

Skills of Interpretation, Inference, and Evaluation

Our students reported high gains in self-efficacy with respect to data analysis, drawing independent conclusions, and evaluating authors' conclusions from papers within or outside their areas of research. However, we observed no statistically significant gains in the postcourse scores in questions assessing the skills of interpretation, inference, and evaluation of data. Possible interpretations of the observed lack of gain in these skills include: limitations of our science-process test, the teaching format in which these skills were practiced, and insufficient amount of time to cause a measurable increase in these skills, as outlined below.

One limitation of our science-process test is that, while it was designed to be accessible to all the students, it was not challenging enough: in most categories of the test, students' mean scores were already high (above 70%) in the pretest. Interestingly, the only component of experimental design that increased significantly, appropriateness of the proposed experiment, was also the most challenging to the students, with the lowest precourse mean score (59.5%). Increasing the difficulty level of our test might increase its sensitivity for measuring these changes over a term. These shortcomings can be addressed in future versions of the science-process skills test.

If the observed difference between students' gains in experimental design but not in interpretation, inference, and evaluation is not due to the limitations of our test, it might be due to the different ways in which these skills were practiced in this course. Students worked in groups to propose three follow-up experiments. Active group participation was encouraged by the use of peer evaluation. On the other hand, the practice of the skills of data interpretation, inference, and evaluation of authors' conclusions was more individualized. Students submitted analyses of three written experiments, typically presented one experiment in front of the class, and listened to their peers presenting the rest of the experiments. Questions to student presenters were encouraged, but typically only several students asked questions after a presentation of each experiment, partly because of time limitations. We believe that it will be of interest in the future to test the effects of additional group activities specifically targeting interpretation, inference, and evaluation skills.

How do our results compare with similar published interventions? Semester-long courses that utilized the CREATE approach to introducing primary literature have been reported to result in statistically significant gains in analyzing data and drawing logical conclusions, as assessed by pre/postcourse tests in an upper-division seminar class (Hoskins *et al.*, 2007), as well as in a freshman-level class, when assessed by the CAT (Gottesman and Hoskins, 2013). It is difficult to draw direct comparison between these studies because of the multiple ways in which they differ: the duration of the course (semester vs. quarter, in our study), the number of papers examined (fewer in our study), the specific format of the course (a sequence of papers from the same lab in CREATE vs. papers from different labs or different fields in our study), and the level of the students (undergraduate vs. master's in this study). Additionally, we designed our test to measure biology content–specific science-process skills, whereas the CAT test (Stein *et al.*, 2012) is not subject specific. Currently, the use of validated critical-thinking tests, such as the CAT (Stein *et al.*, 2012) or the California Critical Thinking Skills Test (Facione and Facione, 1998) is associated with the investment of monetary and faculty time resources (CAT) that were not accessible to the authors. An additional drawback of these tests is that they are not biology specific and therefore do not examine discipline-specific science-process skills.

Are critical-thinking skills discipline specific? We would argue that they have a significant discipline-specific component. Many experts in the critical-thinking field agree that critical thinking within a discipline is deeply connected to discipline-specific knowledge and ways of reasoning (Facione, 1990; Bailin *et al.*, 1999; Willingham, 2008; reviewed in Lai, 2011). For example, in his summary of the Delphi report, Facione states:

> Although the identification and analysis of critical thinking skills transcend, in significant ways, specific subjects or disciplines, learning and applying these skills in many contexts requires domain-specific knowledge. This domain-specific knowledge includes understanding methodological principles and competence to engage in norm-regulated practices that are at the core of reasonable judgments in those specific contexts … Too much of value is lost if critical thinking is conceived of simply as a list of logical operations and domain-specific knowledge is conceived of simply as an aggregation of information. (Facione, 1990, p. 5, quoted in Lai, 2011)

While we think that our test is a useful step in the direction of testing biology-specific critical-thinking and science-process skills, there is a great need for a free, robust, and validated instrument to measure these skills in biology.

Increases in Quantitative Analysis of the Data

We observed a more than twofold increase in the "quantitative data analysis" score, which reflected the frequency and the correctness of quantitative comparisons the students used when describing the differences between experimental and control conditions in the science-process tests. This suggests that the interventions implemented in this course, such as encouraging the use of quantitative description both in written and in oral analyses of experiments, increased students' quantitative literacy, the ability to apply basic quantitative skills to independently analyze data and evaluate claims, an important component of scientific reasoning (National Council on Education and the Disciplines, 2001).

Summary and the Implications of the Increase in Students' Self-Efficacy

How successful was this course in increasing science-process skills of our students? The results we reported point to a mixed success. Using our science-process skills test, we detected statistically significant increases in the use of quantitative terms in data analysis, understanding of controls, and designing an experiment appropriate to the goal of investigation. However, we failed to detect a statistically significant increase in students' performance in questions that required data interpretation, inference, and evaluation of hypotheses. We are currently unable to distinguish between the influences of the course format, the duration of the instruction, and a possible lack of discrimination power of our test as potential interpretations of these findings. A validated and freely accessible instrument that measures students' science-process skills in a biological context will be extremely valuable in assessing the efficacy of different instructional approaches in fostering these skills.

Finally, the instructional approaches described here resulted in significant increases in students' self-efficacy in a variety of science-process skills associated with thoughtful engagement with primary literature, such as independently drawing conclusions from papers' data, evaluating authors' conclusions, and designing a follow-up experiment. The discrepancy between students' self-efficacy and actual academic performance has been previously described in multiple studies (Boud and Falchikov, 1989; Falchikov and Boud, 1989; Dunning *et al.*, 2003, Lawson *et al.*, 2007). Overestimation of academic skills tends to be more pronounced among novices and lower-performing students (Dunning *et al.*, 2003). Because of their graduate status and demonstrated good academic achievement (a requirement for admission into the master's program), our students would be predicted to provide a more accurate assessment of their self-efficacy. Indeed, in the beginning of the course, our students rated their self-efficacy in science-process skills quite low. After instruction, students reported high gains in their self-efficacy, but only limited increases in science-process skills were observed using our test. In future studies, it will be interesting to compare students' self-efficacy and their objective gains in science-process skills using a more robust and validated test.

Collaborative learning and question-and-answer activities, frequently used in this course, are likely contributors to students' gains in self-efficacy. A significant positive effect of these active-learning approaches has been observed in introductory physics classes for nonmajors (Fencl and Scheel, 2005). Personal successful performance of a task is one of the most significant contributors to self-efficacy (Bandura, 1977); therefore, repeated practice of critical analysis of scientific papers in written assignments and oral presentations and the group experimental design activities that were part of this course are also very likely contributors to the observed gains in students' self-efficacy. Observing successful performance of the assessed skills by peers (vicarious experience) is another important contributor to self-efficacy (Bandura, 1977); therefore, multiple student presentations in this course could also have contributed to the overall increase in students' self-efficacy.

The increased self-efficacy is likely to empower students to read more scientific papers and to do so with thoughtful and critical attitudes. It can also have broader effects: a large body of research in psychology documents a positive relationship between academic self-efficacy and students' persistence, performance, and career aspirations (reviewed in Bong and Skaalvik, 2003). For example, Lent and colleagues (1986) showed that undergraduate students who reported higher academic self-efficacy were more likely, 1 yr later, to take courses in science and technology, achieve higher grades in these courses, and consider a wider range of career options in science and technology (this correlation was independent of past academic achievement and interest in the subject). The fact that a 10-wk-long structured engagement with primary literature can produce significant shifts in students' self-efficacy is very encouraging. Further research is required to determine whether the observed increases in self-efficacy correlate with higher performance in following graduate courses, higher number of publications by the students, and metrics of career success.

ACKNOWLEDGMENTS

We are grateful to the biology master's students for their participation in this study; Paula Lemons, Brian Couch, and Alison Crowe for their evaluation and comments on the science-process test; and Sally Hoskins for sharing the materials of the follow-up experimental design assignments ("grant panels," Hoskins *et al.*, 2007). We thank Aaron Coleman, Stanley Lo, Alison Crowe, and our reviewers for their helpful comments. This work was supported by a UCSD Academic Senate Research grant and Biology Scholars Program Alumni Fellowship to E.T.

REFERENCES

Alberts B (2009). Redefining science education. Science *323*, 437.

American Association for the Advancement of Science (2011). Vision and Change in Undergraduate Biology Education: A Call to Action. http://visionandchange.org/files/2011/03/Revised-Vision-and-Change-Final-Report.pdf (accessed 11 September 2014).

American Association of Medical Colleges and Howard Hughes Medical Institute (2009). Scientific Foundations for Future Physicians. www.aamc.org/download/271072/data/scientificfoundationsforfuturephysicians.pdf (accessed 24 September 2014).

Autor DH, Levy F, Murnane RJ (2003). The skill content of recent technological change: an empirical exploration. Q J Econ *118*, 1279–1333.

Autor DH, Price B (2013). The changing task composition of the US labor market: an update of Autor, Levy, and Murnane (2003). Unpublished manuscript. http://economics.mit.edu/files/9758 (accessed 28 September 2014).

Bailin S, Case R, Coombs JR, Daniels LB (1999). Conceptualizing critical thinking. J Curr Stud *31*, 285–302.

Baldwin JA, Ebert-May D, Burns DJ (1999). The development of a college biology self-efficacy instrument for nonmajors. Sci Educ *3*, 397–408.

Bandura A (1977). Self-efficacy: toward a unifying theory of behavioral change. Psychol Rev *84*, 191–215.

Bissell AN, Lemons PP (2006). A new method for assessing critical thinking in the classroom. BioScience *56*, 66–72.

Bloom BS, Krathwohl DR, Masia BB (1956). Taxonomy of Educational Objectives: The Classification of Educational Goals, New York: McKay.

Bong M, Skaalvik EM (2003). Academic self-concept and self-efficacy: how different are they really? Educ Psychol Rev *15*, 1–40.

Boud D, Falchikov N (1989). Quantitative studies of student self-assessment in higher education: a critical analysis of findings. High Educ *18*, 529–549.

Brownell SE, Wenderoth MP, Theobald R, Okoroafor N, Koval M, Freeman S, Walcher-Chevillet CL, Crowe AJ (2014). How students think about experimental design: novel conceptions revealed by in-class activities. BioScience *64*, 125–137.

Coil D, Wenderoth MP, Cunningham M, Dirks C (2010). Teaching the process of science: faculty perceptions and an effective methodology. CBE Life Sci Educ *9*, 524–535.

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

Dasgupta AP, Anderson TR, Pelaez N (2014). Development and validation of a rubric for diagnosing students' experimental design knowledge and difficulties. CBE Life Sci Educ *13*, 265–284.

Dirks C, Cunningham M (2006). Enhancing diversity in science: is teaching science process skills the answer? Cell Biol Educ *5*, 218–226.

Dunning D, Johnson K, Ehrlinger J, Kruger J (2003). Why people fail to recognize their own incompetence. Curr Dir Psychol *12*, 83–87.

Facione PA (1990). Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction. Executive Summary of "The Delphi Report." https://assessment.trinity.duke.edu/documents/Delphi_Report.pdf (accessed 1 April 2015).

Facione PA, Facione NC (1998). California Critical Thinking Skills Test: Form A & B Test Manual, Millbrae, CA: California Academic Press.

Falchikov N, Boud D (1989). Student self-assessment in higher education: a meta-analysis. Rev Educ Res *59*, 395–430.

Fencl H, Scheel K (2005). Engaging students: an examination of the effects of teaching strategies on self-efficacy and course climate in a nonmajors physics course. J Coll Sci Teach *35*, 20.

Gottesman AJ, Hoskins SG (2013). CREATE cornerstone: introduction to scientific thinking, a new course for STEM-interested freshmen, demystifies scientific thinking through analysis of scientific literature. CBE Life Sci Educ *12*, 59–72.

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

Hoskins SG, Stevens LM, Nehm RH (2007). Selective use of the primary literature transforms the classroom into a virtual laboratory. Genetics *176*, 1381–1389.

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

Lai ER (2011). Critical thinking: a literature review. Pearson's Res Rep *6*, 40–41.

Lawson AE, Banks DL, Logvin M (2007). Self efficacy, reasoning ability, and achievement in college biology. J Res Sci Teach *44*, 706–724.

Lent RW, Brown SD, Larkin KC (1986). Self-efficacy in the prediction of academic performance and perceived career options. J Couns Psychol *33*, 265–269.

Muench SB (2000). Choosing primary literature in biology to achieve specific educational goals. J Coll Sci Teach *29*, 255–260.

National Council on Education and the Disciplines (2001). Mathematics and Democracy. The Case for Quantitative Literacy. www.maa.org/sites/default/files/pdf/QL/MathAndDemocracy.pdf (accessed 11 September 2014).

National Research Council (2009). A New Biology for the 21st Century, Washington, DC: National Academies Press.

Paul RW, Elder L, Bartell T (1997). California Teacher Preparation for Instruction in Critical Thinking: Research Findings and Policy Recommendations, Santa Rosa, CA: Foundation of Critical Thinking.

Segura-Totten M, Dalman N (2013). The CREATE method does not result in greater gains in critical thinking than a more traditional method of analyzing the primary literature. J Microbiol Biol Educ *14*, 166–175.

Shi J, Power JM, Klymkowsky MW (2011). Revealing student thinking about experimental design and the roles of control experiments. Int J Sch Teach Learn *5*(2), 1–16.

Sirum K, Humburg J (2011). The Experimental Design Ability Test (EDAT). Bioscene: J Coll Biol Teach *37*, 8–16.

Stein B, Haynes A, Redding M (2012). Critical Thinking Assessment Test, version 5, Cookeville: Center for Assessment and Improvement of Learning, Tennessee Tech University.

University of British Columbia (2014). Questions for Biology (Q4B) Concept Inventories, Experimental Design (Third/Fourth Year Undergraduate Level). http://q4b.biology.ubc.ca/concept-inventories (accessed 20 October 2014).

Willingham DT (2008). Critical thinking: why is it so hard to teach? Arts Educ Policy Rev *109*(4), 21–32.

Zoller U, Lubezky A, Nakhleh MB, Tessier B, Dori YJ (1995). Success on algorithmic and LOCS vs. conceptual chemistry exam questions. J Chem Educ *72*, 987.