Research Article

JMBE Journal of Microbiology & Biology Education

An Active Learning Intervention Based on Evaluating Alternative Hypotheses Increases Scientific Literacy of Controlled Experiments in Introductory Biology

Scott A. Kreher,[a] Iglika V. Pavlova,[b] and April Nelms[c]
[a]*Dominican University, River Forest, Illinois, USA*
[b]*University of North Carolina Greensboro, Greensboro, North Carolina, USA*
[c]*University of North Georgia, Dahlonega, Georgia, USA*

Scott A. Kreher and Iglika V. Pavlova contributed equally to this article. Author order was determined alphabetically.

Scientific education provides a set of tools to make sense of a complex world by teasing out complicated cause-and-effect relationships, such as the elimination of effects of confounding factors in controlled experiments. There is evidence that depth of understanding of controlled experiments is lacking among undergraduate science students despite exposure to controlled experiments in courses. To examine the understanding of controlled experiments, we developed a two-tiered assessment that includes closed-ended and open-ended questions, with three types of questions, i.e., (i) a scientific scenario about a flawed drug study, (ii) an everyday-life scenario about flawed thinking regarding product effectiveness, and (iii) a direct question about explaining controlled experiments. Consistent with previous findings, we demonstrated that large percentages of students in introductory biology courses at both a research-intensive institution and a primarily undergraduate, minority-serving institution failed to recognize the need to account for confounds. Based on these findings, we tested the hypothesis that scientific literacy could be improved through a course-based intervention using an active learning framework focused on science as a process of evaluating alternative hypotheses. We found start-to-end-of-semester improvement in students' identification of unaccounted confounds with a scientific scenario in an intervention course but not in the control course. Interestingly, students in both the control and intervention courses showed improvement when tested with a scenario based on everyday life. The study findings suggest that a concerning number of college students may not account sufficiently for uncontrolled variables in real-life situations, and we present a widely applicable instructional strategy that improves on this broadly relevant scientific reasoning skill.

KEYWORDS scientific literacy, controlled experiments, active learning, alternative hypotheses

INTRODUCTION

The *Vision and Change* report identifies the skill of applying the process of science as a core competency in the undergraduate biology curriculum, toward the goal of lifelong scientific literacy (1). Science as a process can be expanded into multiple realms, including scientific inquiry (SI) and the nature of science (NOS) (2), both of which are emphasized as learning goals in the U.S. Next Generation Science Standards (NGSS) for primary and secondary education (https://www.nextgenscience.org). A major component of SI is to investigate the complex phenomena of our world by determining cause-and-effect relationships; underlying this exploration is accounting for confounding variables that impact the effect of interest, which is often accomplished through controlled experiments (2). There is value in teaching the concepts of confounding factors, testing alternative hypotheses, and controlled experiments as epistemic components of the NOS, which may lead to stronger intellectual foundations (3, 4). A value of the emphasis on explicit teaching of controlled experiments is that we can help students understand SI and the NOS while also promoting critical thinking.

While it is a reasonable assumption that students develop a depth of understanding of controls and controlled experiments due to exposure in laboratory courses or during research experiences, multiple studies indicate that

Citation Kreher SA, Pavlova IV, Nelms A. 2021. An active learning intervention based on evaluating alternative hypotheses increases scientific literacy of controlled experiments in introductory biology. J Microbiol Biol Educ 22:e00172-21. https://doi.org/10.1128/jmbe.00172-21.
Address correspondence to Dominican University, River Forest, Illinois, USA. E-mail: skreher@dom.edu.
Received: 2 June 2021, Accepted: 27 September 2021, Published: 29 October 2021

students have crucial gaps in their understanding when explicit teaching is lacking. For example, in an observation of secondary-education-level classes in which students were involved in inquiry-based activities, students did not discuss or identify controls in their experimental planning (5). Shi and colleagues found that undergraduate students could not adequately explain experimental controls even after completing three laboratory courses (6). Other studies confirmed that exposure to controlled experiments is not sufficient to lead to a depth of understanding of controls (7–10).

Assessment of understanding of the scientific process is not trivial. Whereas a number of existing assessments for scientific literacy and experimental design skills have revealed changes in student literacy, these assessments use a multiple-choice format (https://www.aaas.org/programs/project-2061) (11, 12) and do not directly interrogate the process of student reasoning. Other instruments evaluate students' skills in designing an experiment but do not necessarily require explanation of the process or the depth of understanding (13–15). The existing instruments do not necessarily aim to measure skills in real-life contexts where there may be distractors or where cues that a student is supposed to design a scientific experiment may be lacking. The development of assessments that have open-ended questions, contain distractors, or simulate real-life situations may be valuable in probing for the transfer of science process skills to real-life contexts.

Another difficulty in teaching the process of science stems from the longstanding tension in the perceived trade-off between content and science process skills/critical thinking, especially for lecture courses in the content-heavy science disciplines. In a survey of undergraduate faculty, most were strongly supportive of teaching science process skills but often did not incorporate them into lecture courses due to the pressure to cover content (16). Biology lecture curricula that successfully integrate science process skills to support the teaching of content may better promote the development of useful scientific literacy (17–19).

Given the complexity of the process of science, scientific literacy, and challenges in explicit teaching of experimental controls and confounding factors in already overfull curricula, we sought to create a flexible framework as an intervention that was grounded in the philosophy of science. We chose an overall active learning framework of science as a process of evaluating two or more competing hypotheses (3). The importance of evaluating multiple, alternative hypotheses has long been recognized as a cause of scientific progress (20), and many conceptualizations of the NOS include methods of evaluating alternative hypotheses (21). Hastie and Dawes wrote that scientific reasoning is defined as examining alternative explanations using evidence (22). While not every aspect of biology can fit into the framework of comparing alternative hypotheses, there are many good examples, both historical (is the genetic material DNA or protein?) and more recent (did severe acute respiratory syndrome coronavirus 2 [SARS-CoV-2] move from bats to humans directly or did the virus move from bats to a secondary host and then to humans? [23]). There is also a developing recognition of the importance of consideration of alternative hypotheses by practicing scientists and critical self-reflection regarding why some disciplines should more strongly emphasize this component of the NOS (24).

As part of our framework, controlled experiments can be seen as special cases of eliminating alternative hypotheses and explanations. We emphasized that confounding factors could be conceptualized as alternative explanations or hypotheses that confuse cause-and-effect relationships between independent and dependent variables and that the purpose of controlled experiments is to eliminate or account for the effects of confounding factors. The value of the intervention is that it can complement associated laboratory courses, and specific examples can be tailored to instructor interest to support the teaching of content, rather than detracting from it. To further support the teaching of content in lecture courses, especially in cases where the course does not have an associated laboratory component, there is explicit instruction on the broader reasons to consider confounding factors, which can be taught within content domains of undergraduate introductory biology, such as evolution, cell biology, and energy transformation (1).

In order to assess the effectiveness of our framework, we first sought to examine the depth of understanding of science as a process over the course of undergraduate introductory biology courses in two different types of institutions, using a two-tiered assessment (described in Procedure). Hypothesis 1 is that students would have difficulty identifying that confounding factors were unaccounted for when faced with a flawed scientific scenario and there would not be improvement throughout the course. The main evidence underlying the expected outcome of hypothesis 1 is the finding of multiple previous studies in which student understanding of controlled experiments did not improve after exposure to controlled experiments in undergraduate courses (6–10). After establishing the wide prevalence of this problem, we taught an introductory biology course that used (i) the framework centered on thinking about alternative hypotheses and confounding factors when determining cause-and-effect relationships together and (ii) multiple opportunities to practice application using case studies fully integrated with course content in an introductory biology course. Hypothesis 2 is that students in the intervention course would show improvement in identifying confounding factors, compared to students in the control group, which also used an active learning method but did not present the generalizable framework, as assessed using our two-tiered assessment with a flawed scientific scenario and a flawed everyday-life scenario. The reasoning behind the predicted outcome for hypothesis 2 is that successful learning due to the framework should be measured as application of concepts fundamental to the NOS. We found that students in the intervention introductory biology course did have improved explanations of confounding factors and alternative hypotheses in response to a scientific scenario, but both control and intervention courses improved in response to a scenario grounded in everyday life.

TABLE 1
Summary of our two-tiered assessment process and overview of our intervention[a]

Scenario	Assessment questions	
	Tier 1: closed-ended	**Tier 2: open-ended**
Scientific scenario	Is a drug effective? (in a scenario lacking controls) (effective, somewhat effective, don't know, somewhat ineffective, or ineffective)	Is a drug effective? Explain your answer in writing:
Everyday-life scenario	Is a product effective? (in a scenario lacking controls) (effective, somewhat effective, don't know, somewhat ineffective, or ineffective)	Is a product effective? Explain your answer in writing:
What is a control?		What is a control or controlled experiment? Explain your answer in writing:

[a]The intervention course was an introductory biology course using a framework of science as a process of testing alternative hypotheses. The control course was an introductory biology course taught by the same professor; it was taught with the same course topics as the intervention course but not within the framework. Assessments were given at the beginning and end of the semester. For more details, see Procedure and the supplemental material.

PROCEDURE

Assessment design and validation

We designed an assessment that prompts responses to three realistic scenarios, namely, scientific, everyday life, and "what is a control?" (Table 1; also see Appendix S1 in the supplemental material). The scenarios test for student understanding of the general issue of the role of confounds in determining cause-and-effect relationships in three different contexts. In the first two scenarios, students are asked whether a certain possible cause of interest (either a drug for the scientific scenario or a product from daily life for the everyday-life scenario) is effective in determining an outcome. The scenarios present situations in which accounting for confounds would be necessary to make a justified conclusion, but they are unaccounted for as described. We employed a two-tiered assessment for the scientific and everyday-life scenarios. The tier 1 assessment included closed-ended answer options regarding whether the cause of interest is effective in determining the outcome (closed-ended answer options included effective, somewhat effective, don't know, somewhat not effective, and not effective); this was followed by open-ended tier 2 "explain your reasoning" responses to the same questions from tier 1. In the third scenario, i.e., what is a control?, students were asked to explain controls, in written form, to a friend at a party.

The assessment prompts underwent several rounds of revision, using feedback from students and instructors to improve validity (see Appendix S1 and Appendix S2 in the supplemental material). First, the assessment underwent rounds of revision and pilot testing with students and instructors in undergraduate biology courses to improve the face validity. Phrasing or prompts that were confusing to students were eliminated. Additional revisions were made based on input from biology and science education faculty. Second, a focus group was conducted at the study site, Dominican University, where we found that students did understand the questions and were able to reason through the scenario components. Third, we compared the results of our assessment with validated AAAS questions on control of variables (https://www.aaas.org/programs/project-2061) and an instrument on scientific literacy, the Test of Scientific Literacy Skills (12). Fourth, we compared independent groups of students tested with our assessment with added questions with scenarios in which confounding factors were controlled, finding that students could accurately make conclusions. The validated assessment was integrated into the curriculum, with the precourse assessment occurring at the start of the semester, presented as a warm-up to scientific thinking, and the postcourse assessment occurring at the end of the semester, presented as a concluding activity in the course.

Participants

Data were collected from undergraduate students enrolled in introductory biology courses at two universities, namely, a selective research institution (R1) (University of Chicago) and an inclusive predominantly undergraduate institution (PUI) (Dominican University); the PUI serves a large number of underrepresented students, predominantly from Latinx backgrounds (61% of undergraduate students are Hispanic/Latinx), and is classified as an Hispanic-serving institution. The R1 has two types of introductory biology courses, i.e., one for nonscience majors and one for majors, which were analyzed separately. The PUI has only one level of introductory biology course, for biology majors and nonmajors together. To reduce bias in the student data set, all student responses were included in the final data sets as long as the students consented and answered both precourse and postcourse survey questions.

The intervention and control occurred in two different sections of the same introductory biology course at the PUI. Students were randomly assigned to the control course or the intervention course at the PUI by registration staff members; students had no knowledge of the intervention prior to the course, and professors had no role in student registration. Both courses were taught by the same professor. Students in the control and intervention courses had similar mean

secondary school grade point averages (GPAs) and similar mean standardized test scores.

Data collection

The researchers collected data in the form of digital assessment responses, which were collected in Learning Management Software in the first 2 weeks of each course (precourse survey) and during the last 2 weeks of each course (postcourse survey). Data were deidentified within course groups, and written tier 2 responses were masked for both student identity and treatment group, to allow blind scoring to reduce possible bias.

Data analysis

Data were analyzed only from students who were over 18 years of age, gave their consent, and completed both the precourse and postcourse surveys. Studies at both universities were deemed exempt under institutional review board protocols RCAS 11–21 (Dominican University) and H09468 (University of Chicago). All data collection, including that from intervention and control courses, occurred over 3 consecutive academic years.

Tier 2 written responses were analyzed by coding. For tier 2 responses, a coding rubric (see Appendix S3 in the supplemental material) was developed through several rounds of coding of a subset of the data and discussion by the authors. To reduce bias, deidentified written responses were blind-coded to consensus by two researchers according to the developed coding rubric; researchers were blinded to student identity and response group (control versus intervention) and were also blinded to tier 1 answers. The "what is a control?" scenario was additionally coded for emerging themes on how students describe controlled experiments. All responses were coded by two authors, with one author serving as a common coder for all data sets to ensure consistency in coding. Interrater reliability was over 80% in all cases.

The closed-ended assessment data and open-ended coded essay data were analyzed for changes from the precourse assessment to the postcourse assessment with McNemar's tests (2 × 2 contingency tables) or Stuart-Maxwell tests (3 × 3 contingency tables). Analysis of possible relationships between closed-ended and open-ended responses was conducted with the nonparametric Kruskal-Wallis test, with *post hoc* Dunn's test of multiple comparisons. Closed-ended responses were coded in three levels, i.e., effective, not effective, and don't know, and composite scores for open-ended responses, ranging from 0 to 3, were calculated for positive scores for three dimensions; a score of 3 means that a student scored positive for all three dimensions. All statistical analyses were performed with R.

Intervention

The intervention implemented in introductory biology courses at the PUI was composed of an active learning intervention using case studies and exercises centered on comparing possible alternative explanations and accounting for confounding variables, in multiple ways and at multiple levels of organization (for more detail, see Appendix S4 and Appendix S5 in the supplemental material). Students were randomly assigned, via registration, into either the intervention or control course, with no prior student knowledge of the intervention; course registration was conducted by university advising staff members without instructor knowledge. The same course topics which are typical topics for a first-semester, undergraduate biology course, were covered in each course (intervention and control). There was no detectable bias in student composition of the intervention or control groups, and students had similar average secondary school grades (measured by GPA) and standardized examination scores.

Instruction in the intervention group used a generalizable framework of scientific thinking that emphasized the importance of considering alternative hypotheses where appropriate (3), and controlled experiments could be seen as a special case of the framework. Active learning activities included in-class group problem-solving and discussion and interactive questions using clicker technology. The control course also received active learning activities, but they did not include the framework of the intervention. The same professor taught the control and intervention courses.

The exercises and case studies of the intervention were composed of student-centered activities focused on major content themes that invited students to consider how alternative hypotheses are useful for understanding a wide range of biology concepts. For example, students practiced designing experiments using "inductive space" methods and drawing on their prior knowledge, such as designing an experiment to test whether or not a person is lactose intolerant (2). Students also used prior knowledge and directed research to explore inductive space regarding key biological phenomena, such as the evolutionary relatedness of whales to other mammals and seemingly similar vertebrates, such as fish. Other activities focused on the role of alternative hypotheses as presented in secondary and tertiary literature (such as articles from *Scientific American*), such as possible evolutionary explanations for why humans are relatively hairless, compared to most mammalian relatives. Scaffolded instruction was used to guide students to read primary literature to introduce cases in which alternative hypotheses were tested and discussed, such as alternative hypotheses of why many human tumor cells enter alternative metabolic states, compared to nontransformed cells. More details and a list of references, example exercises, and introductory biology topics can be found in Appendix S5 in the supplemental material.

RESULTS

Initial scientific scenario assessment

We first examined how the understanding of the need to account for confounding factors changes over the course of undergraduate introductory biology courses at two different

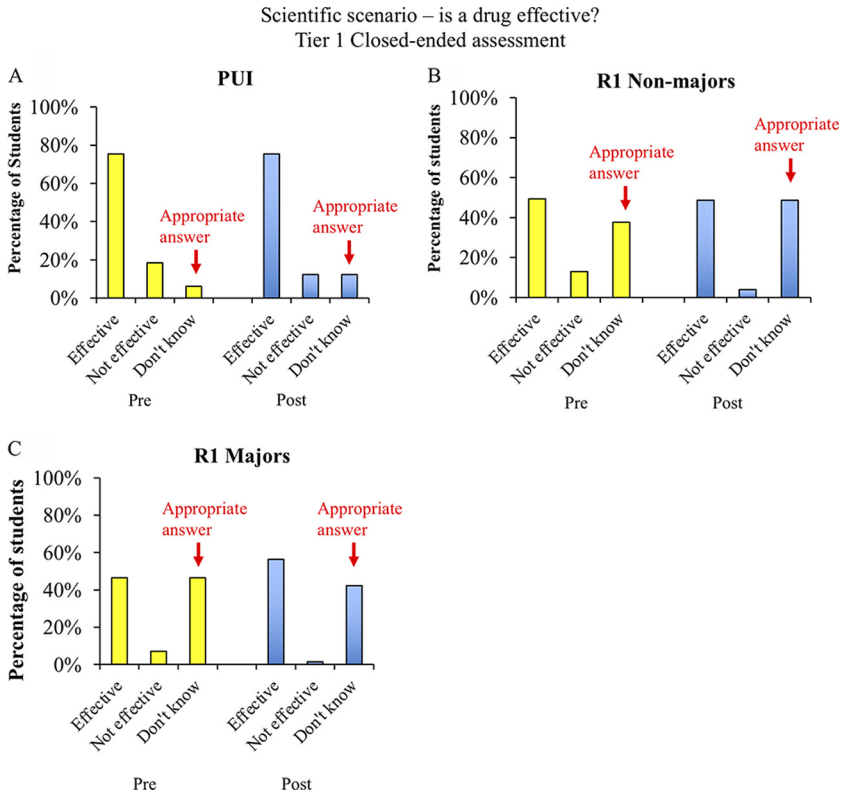

FIG 1. Closed-ended (tier 1) assessment in response to the scientific scenario (is a drug effective?). Students were given a scenario in a precourse/postcourse start-of-semester/end-of-semester format and asked whether the drug was effective. Closed-ended answer options included effective, somewhat effective, don't know, somewhat not effective, and not effective. Effective/somewhat effective and not effective/somewhat not effective answers were pooled. Given the scenario, the don't know answer is most appropriate. (A) Responses of students in an introductory biology course (mixture of biology majors and non-biology majors) at a small liberal arts college/PUI, with a large underrepresented student population ($n = 65$ from two separate courses). (B) Responses of non-biology majors in a nonmajors introductory biology course at a private R1 ($n = 77$). (C) Responses of biology majors in a majors introductory biology course at a private R1 ($n = 71$). No groups showed statistically significant differences from the precourse assessment to the postcourse assessment by the Stuart-Maxwell test ($P > 0.05$).

institutions, using a two-tiered assessment; tier 1 was a closed-ended scientific scenario lacking any mention of experimental controls, and tier 2 was an open-ended assessment that allowed students to write and explain their reasoning in response to the questions from tier 1 (Table 1). Consistent with the findings of Shi and colleagues, in which students had difficulty understanding controlled experiments despite exposure in biology courses (6), we also found that students were generally not able to account for confounding factors and that there was no improvement throughout introductory biology courses.

For the tier 1 scientific scenario, a remarkable 75% of students at the PUI ($n = 65$) erroneously indicated in the precourse assessment that the drug was effective, despite the fact that the scenario lacked any mention of accounting for confounding factors, such as having a control group (Fig. 1A). In contrast, 50% of non-biology majors ($n = 77$) and 47% of biology majors ($n = 71$) at the R1 indicated that the drug was effective (Fig. 1B and C). There was variation in each group's responses for the don't know answer option, which is the most appropriate choice, given the significant issues in experimental design in the presented scenarios. Whereas only 6% of students at the PUI chose don't know, 38% of nonmajors and 47% of majors chose don't know at the R1 university (Fig. 1). The not effective category was chosen by 19% at the PUI, 13% of R1 non-biology majors, and 7% of R1 biology majors.

We also assessed whether answer choices changed over the course of the semester. Instructors in all courses used some level of active learning (in the form of in-class small-group activities) and had a lecture on scientific method; all student populations had a laboratory section associated with the course, in which they performed scientific experiments. Overall, there was no significant precourse-to-postcourse change over the course of the semester in the percentage of students, among PUI students, R1 nonmajors, or R1 majors, who erroneously stated that the drug was effective (Stuart-Maxwell test, $P > 0.05$) (Fig. 1).

Evidence from our assessment validation from an independent group of students supports the finding that students can accurately choose tier 1 answers when presented with two scenarios that do account for confounding factors; 93% of students accurately chose effective in a scenario in which there was an effect, and 75% of students accurately chose not effective when there was not an effect (see Appendix S1 in the supplemental material).

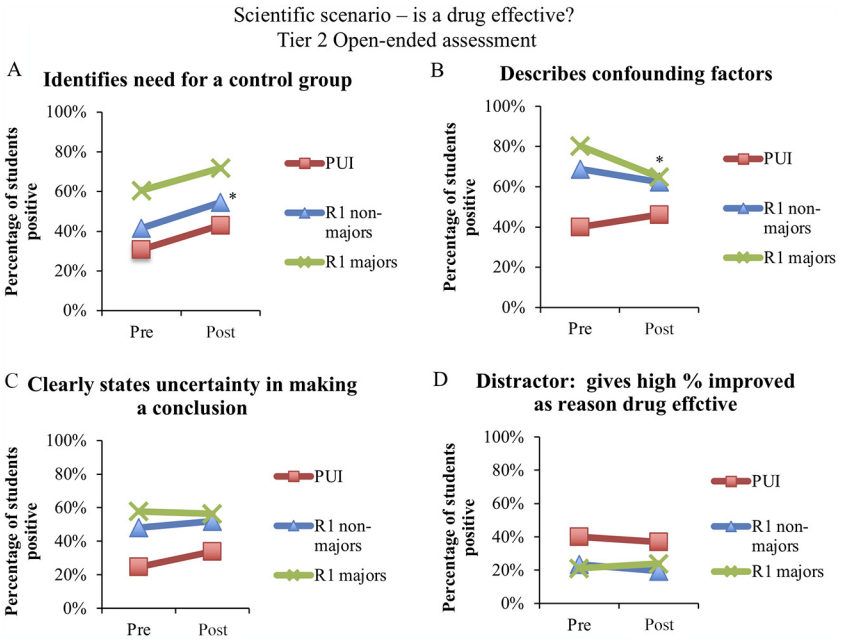

FIG 2. Open-ended (tier 2) assessment in response to the scientific scenario (is a drug effective?). The same students who answered the close-ended assessment described in Fig. 1 explained their tier 1 answers in written form. Written responses were scored along four dimensions (A to D) for each student population. Asterisks indicate significant differences by McNemar's test ($P < 0.05$).

We then sought to examine student reasoning for student tier 1 answers through coding of tier 2 responses. For the first dimension, only 26% of PUI students mentioned the lack of controls, whereas 41% and 61% of R1 nonmajors and biology majors, respectively, did so (Fig. 2A). For the second dimension, 40% of PUI students, 69% of R1 nonmajors, and 80% of R1 majors successfully identified one or more confounding factors that would prevent reaching a conclusion from the scenario (Fig. 2B). The third dimension, i.e., statement of uncertainty in making a conclusion, was scored by noting direct statements by students such as "I cannot make a conclusion here" or "There is no way to know, given the information." A total of 25% of PUI students scored positive for this dimension, whereas 48% of nonmajors and 58% of majors at the R1 scored positive (Fig. 2C). For the fourth dimension, i.e., whether or not numerical distractor data were given as evidence supporting a conclusion, 40% of PUI students, 23% of R1 nonmajors, and 20% of R1 majors scored positive (Fig. 2D).

Similar to tier 1 answers, tier 2 scores were fairly stable over the semester. Average scores for written responses from the postcourse assessment were typically within 10 percent of the precourse assessment scores. There was one statistical increase from the precourse assessment to the postcourse assessment, namely, responses of R1 nonmajor students on the dimension of identifying the need for a control (McNemar's test for paired data, chi-square = 4.05, df = 1, P = 0.04) (Fig. 2A). There was one statistical decrease, namely, responses of R1 majors on the dimension of describing confounding factors (McNemar's test for paired data, chi-square = 5.88, df = 1, P = 0.02) (Fig. 2B); we note that the precourse assessment scores for this group were very high and among the highest of any dimension.

Scientific scenario in intervention and control groups

We designed and implemented a course-level intervention in an introductory biology course at the PUI based on consideration of alternative hypotheses as an aspect of the NOS (described in Procedure; also see Appendix S4 and Appendix S5 in the supplemental material). We found a significant intervention effect when comparing the intervention (n = 39) and control (n = 36) groups with respect to the scientific scenario closed-ended tier 1 assessment (Fig. 3). Whereas most students in the control group chose the erroneous answer effective in both precourse and postcourse tier 1 assessments (Stuart-Maxwell test statistic = 2.5, df = 2, P = 0.29), there were significant shifts in the responses of the intervention group, with 28% of students choosing don't know and 38% choosing not effective answers in the postcourse assessment (Stuart-Maxwell test statistic = 14.4, df = 2, $P < 0.01$). For tier 2 written responses, students in the control group did not show any significant changes in any of the four dimensions ($P > 0.05$) (Fig. 4A to D). In contrast, students in the intervention group improved, with significant precourse-to-postcourse changes in three of the four dimensions (Fig. 4B to D). Specifically, the proportion of students describing confounding factors (intervention group, McNemar's test for paired data, chi-square = 6.05, df = 1, P = 0.01) (Fig. 4B) and clearly stating uncertainty in making conclusions (intervention group, McNemar's test for paired data, chi-square = 6.26, df = 1, P = 0.01) (Fig. 4C) increased significantly from the precourse assessment to the postcourse assessment, while there was a significant decline in the proportion of students who cited the quantitative distractor data (intervention group, McNemar's test for paired data, chi-square = 4.92, df = 1, P = 0.03) (Fig. 4D). Overall,

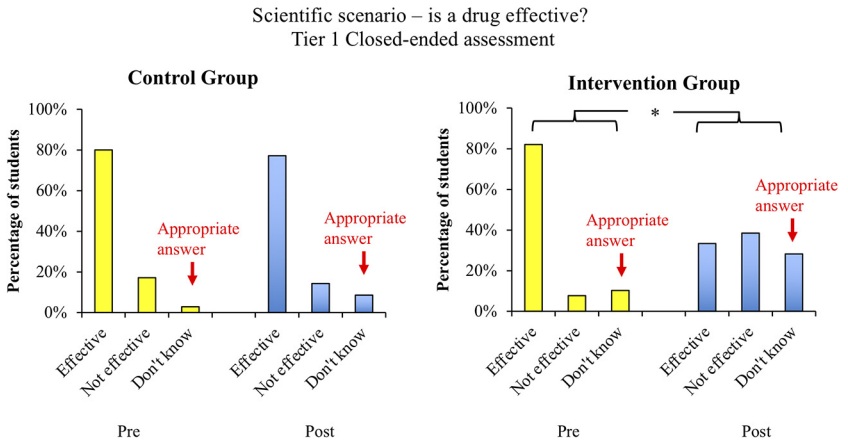

FIG 3. Closed-ended (tier 1) assessment in control and intervention groups in response to the scientific scenario (is a drug effective?). Closed-ended answer options included effective, somewhat effective, don't know, somewhat not effective, and not effective. Effective/somewhat effective and not effective/somewhat not effective answers were pooled. Given the scenario, the don't know answer is most appropriate. Two introductory biology courses at the PUI, as described for Fig. 1 and 2, were administered the control or intervention active learning curriculum. The control course is a single course from the PUI data in Fig. 1 and 2 ($n = 36$); the intervention course is a separate course taught by the same instructor ($n = 39$). The intervention group but not the control group showed significant precourse to postcourse changes by the Stuart-Maxwell test ($P < 0.05$), designated by an asterisk.

students in the intervention group for the scientific scenario improved significantly over those in the control group, as evidenced by the tier 1and tier 2 assessments.

We found that tier 2 answers corroborated tier 1 answers with the nonparametric Kruskal-Wallis tests, using tier 1 categories and tier 2 composite essay scores ("effective" mean, 1.18; "don't know" mean, 2.04; Kruskal-Wallis test, chi-square = 25.1, df = 2, $P = 0$). From these results, we conclude that closed-ended answers of don't know are associated with more accurate and higher-quality open-ended responses, indicating that data from tier 1 and tier 2 answers are indicating similar views of student understanding.

Everyday-life scenario in intervention and control groups

We also examined how well the intervention would allow transfer to a different kind of scenario that was drawn

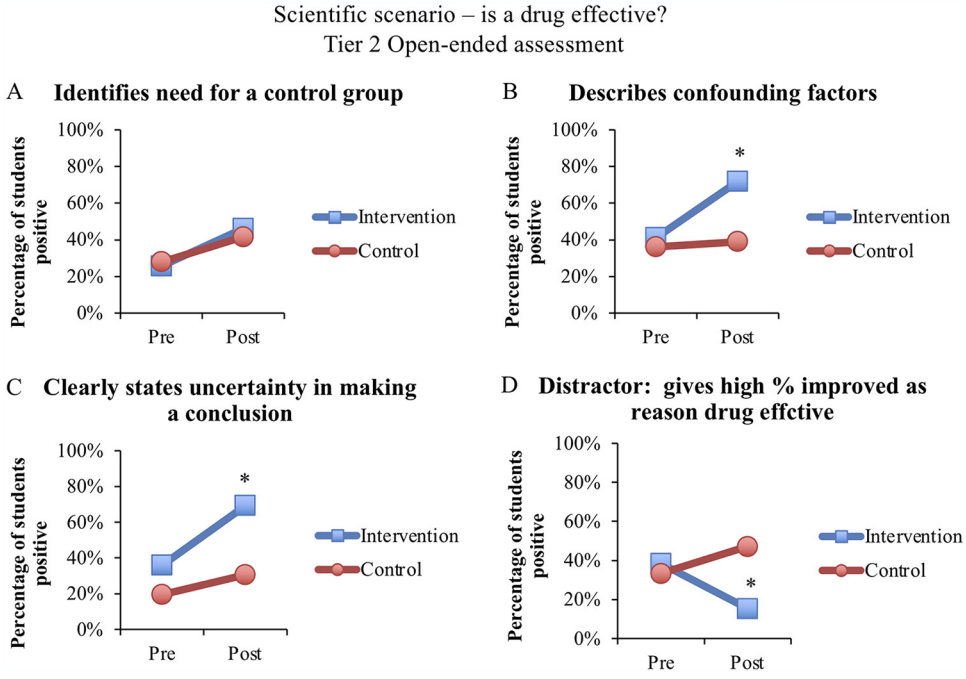

FIG 4. Open-ended (tier 2) assessment in the control and intervention groups in response to the scientific scenario (is a drug effective?). The control and intervention groups are the same groups as described for Fig. 3 ($n = 36$ and $n = 39$, respectively). Written responses were scored along four dimensions (A to D) for each student population. Asterisks indicate significant precourse-to-postcourse differences by McNemar's test ($P < 0.05$).

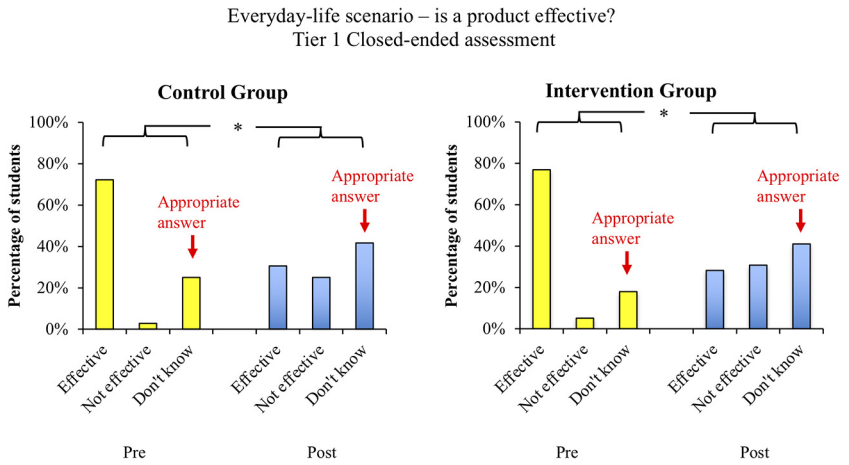

FIG 5. Closed-ended (tier 1) assessment in the control and intervention groups in response to the everyday-life scenario (is a product effective?). Closed-ended answer options included effective, somewhat effective, don't know, somewhat not effective, and not effective. Effective/somewhat effective and not effective/somewhat not effective answers were pooled. Given the scenario, the don't know answer is most appropriate. The control and intervention groups are the same groups as described for Fig. 3 and 4 ($n = 36$ and $n = 39$, respectively). The intervention and control groups both showed significant precourse-to-postcourse changes by the Stuart-Maxwell test ($P < 0.05$), designated by asterisks.

from everyday life and does not have the appearance of a scientific study. Using daily-life examples on how to improve cooking and stain removal, we asked whether products were effective despite their lack of accounting for confounding factors or a control. Interestingly, we found significant precourse-to-postcourse changes in both the control and intervention groups, with more students choosing don't know or not effective in the postcourse assessment, compared to the precourse tier 1 assessment (control group, Stuart-Maxwell statistic = 11, df = 2, $P < 0.01$; intervention group, Stuart-Maxwell statistic = 14.8, df = 2, $P < 0.01$) (Fig. 5).

In the analysis of the tier 2 written responses for the everyday-life scenario, both control and intervention groups showed improvement in three of four dimensions, with significant increases from the precourse assessment to the postcourse assessment in two dimensions shared by the two groups (control and intervention), i.e., describing confounding factors (control group, McNemar's test for paired data, chi-square = 15.06, df = 1, $P < 0.01$; intervention group,

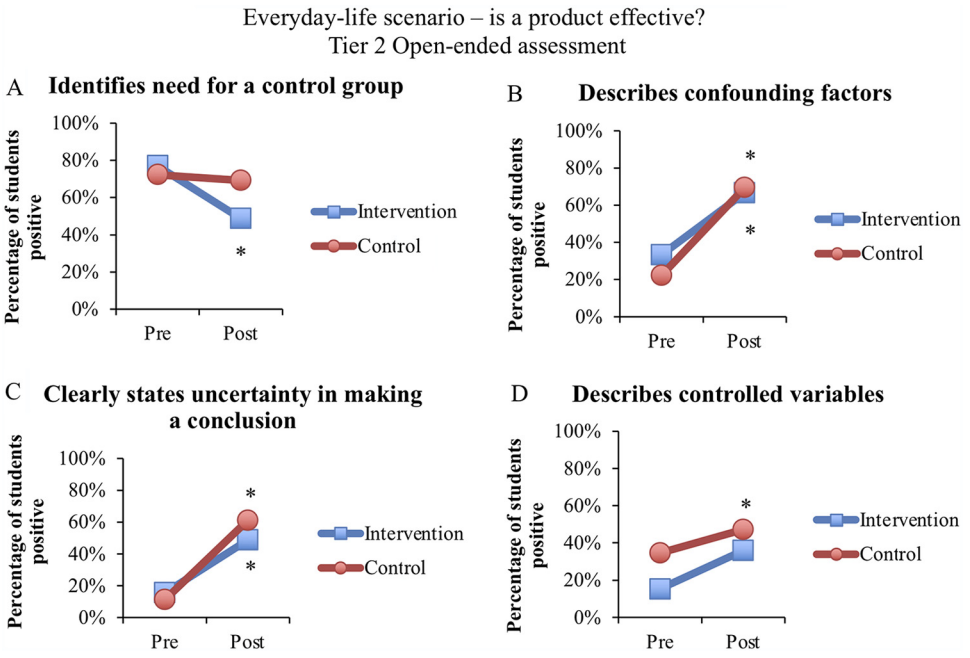

FIG 6. Open-ended (tier 2) assessment in the control and intervention groups in response to the everyday-life scenario (is a product effective?). The control and intervention groups are the same groups as described for Fig. 3 to 5 ($n = 36$ and $n = 39$, respectively). Written responses were scored along four dimensions (A to D) for each student population. Asterisks indicate significant differences from the precourse assessment to the postcourse assessment by McNemar's test ($P < 0.05$).

McNemar's test for paired data, chi-square = 5.33, df = 1, $P = 0.02$) (Fig. 6B) and clearly stating uncertainty in making conclusions (control group, McNemar's test for paired data, chi-square = 14.45, df = 1, $P < 0.01$; intervention group, McNemar's test for paired data, chi-square = 7.58, df = 1, $P < 0.01$) (Fig. 6C), and a third dimension for the control group alone, i.e., describing controlled variables (control group, McNemar's test for paired data, chi-square = 4.9, df = 1, $P = 0.03$) (Fig. 6D). Interestingly, the intervention group showed a significant decrease in the proportion of students who stated that a control was necessary to make a conclusion (intervention group, McNemar's test for paired data, chi-square = 4.35, df = 1, $P = 0.04$) (Fig. 6A). Overall, for the everyday-life scenario, students improved over the course of the semester in both the control and intervention groups.

What is a control? scenario in intervention and control groups

In the third assessment prompt, the "what is a control?" scenario, students were asked to explain controls to a friend at a party. The what is a control? scenario included only written responses. Open-themed coding for patterns in student responses revealed similar prevalences of three major emergent themes among the control and intervention groups in the precourse assessment (Table 2). Most students described controls as comparisons (control group, 67%; intervention group, 74%), followed by descriptions of controls as knowns (control group, 47%; intervention group, 36%) or as something that does not change (control group, 22%; intervention group, 28%). Over one-half of the students used two or three of these ideas in their explanation. At the end of the semester, the same three emergent themes dominated student answers in the control course, with an increase toward controls being described as knowns (64%) and a decrease in controls being described as comparisons (44%). In contrast, in the intervention group, there was a shift in the postcourse assessment toward a new theme, i.e., controls eliminating alternative explanations, being the most prevalent in 67% of postcourse student answers (precourse, 0%). Controls being explained as comparisons increased (89%) in the postcourse assessment.

Answers for the "what is a control?" scenario were also coded for three dimensions aimed at assessing the quality of explanations, specifically, whether students (i) describe how controls allow conclusions regarding cause-and-effect relationships, (ii) explain how controls account for confounding factors, and (iii) describe clearly comparing control and experimental groups. We note that, even in the precourse survey, up to one-half of the students scored positively for describing cause-and-effect relationships in their responses. However, there were no significant improvements in the quality of explanations from the precourse survey to the postcourse survey in either the control or intervention group (McNemar's test for multiple comparison, $P > 0.05$) (Fig. 7).

TABLE 2
Open-ended assessment using emergent theme coding in control and intervention groups in response to the prompt: what is a control?[a]

Emergent theme and group	Percentage of students who scored positive	
	Precourse	Postcourse
Controls are comparisons		
Control	66.7%	43.6%
Intervention	74.4%	88.9%
Controls are knowns		
Control	47.2%	63.9%
Intervention	35.9%	33.3%
Controls are constant/do not change		
Control	22.2%	11.1%
Intervention	28.2%	23.1%
Controls eliminate alternative explanations		
Control	2.8%	5.6%
Intervention	0.0%	66.7%

[a]The control and intervention groups are the same groups as described for Fig. 3 to 7 ($n = 36$ and $n = 39$, respectively). Written responses were coded for emergent themes, with the four most prevalent themes being presented.

DISCUSSION

Our assessment revealed that large proportions of students in two institutions with very different student bodies draw erroneous conclusions despite the lack of stated methods to account for confounding factors in flawed scientific scenarios, supporting our first hypothesis. Even in the better-performing group from the selective R1, up to one-half of the student answers did not express appropriate skepticism when it was warranted from the scenario (Fig. 1 and 2). While we observed variation in responses across two university populations, there was little improvement in any student population over a biology course in the absence of an intervention. Despite the widely recognized importance of learning the process of science, such as concepts of controlled experiments, confounding factors, and alternative explanations, explicit teaching is often lacking in undergraduate curricula (16). An assumption that students will learn about the process of science through just exposure via teaching labs is not borne out by the evidence, and our study confirms the previous results (6).

Our study demonstrates that the intervention framework of teaching biology as a process of considering alternative hypotheses can increase the depth of understanding of certain aspects of the process of science, partially supporting our second hypothesis. While our intervention led to improved scientific literacy when interrogated with the scientific scenario, compared to the control course (Fig. 3 and

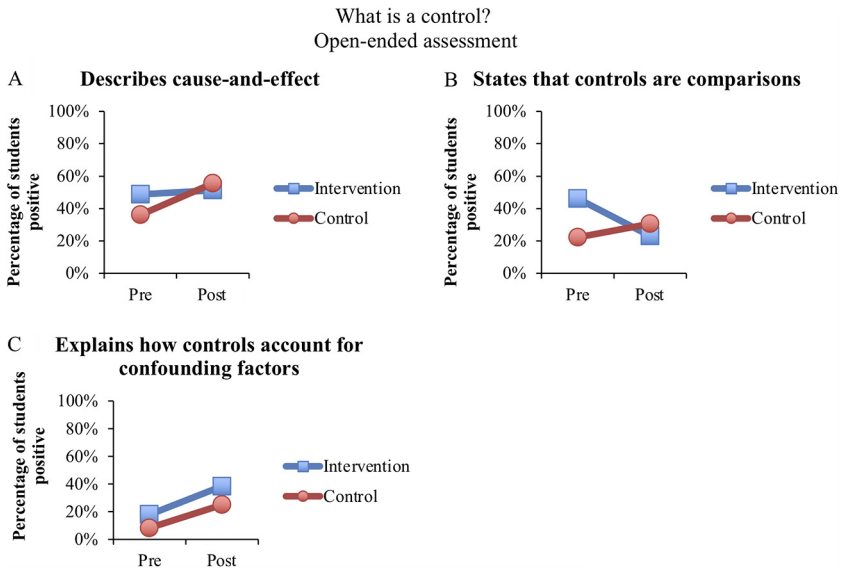

FIG 7. Open-ended assessment in the control and intervention groups in response to the prompt: what is a control? The control and intervention groups are the same groups as described for Fig. 3 and 6 ($n = 36$ and $n = 39$, respectively). Written responses were scored along three dimensions (A to C) for each student population. No groups showed statistically significant differences from the precourse assessment to the postcourse assessment by McNemar's test ($P > 0.05$).

4), we found that scientific literacy improved in all students when interrogated with the everyday-life scenario (Fig. 5 and 6). Interestingly, responses to a question that prompted more direct explanations of experimental controls (what is a control?) did not significantly improve in either the intervention group or the control group (Fig. 7 and Table 2).

Other researchers have found that, in the absence of explicit teaching about the epistemic basis of controlled experiments, students do not show improved understanding after completing laboratory courses in which students are exposed to controlled experiments (6, 8, 10), and our results confirm this growing body of evidence. However, while previous researchers found that scientific literacy could be improved by exercises on experimental design, the interventions were usually limited to course-based labs and assessed experimental design skills in a limited manner (6, 11–15, 25). While experimental design skills are important, especially for science majors, improved scientific literacy is a goal for all students. Our work adds to the field by providing an intervention that can be used in lecture courses, which could provide a way to improve scientific literacy among undergraduates who are not science majors and may not take a laboratory course. Our assessment also provides a tool that can assess the understanding of controlled experiments through analysis of writing, which could reveal conceptual understanding. An interesting question that emerges from our work is whether students have improved experimental design skills after the intervention. While our intervention does have some conceptual exercises related to experimental design, future work should examine how better conceptual understanding may transfer to specific experimental design skills.

Another interesting finding of our study is that students improved in their understanding of confounding factors in both the intervention and control courses in response to the scenario grounded in everyday life (Fig. 5 and 6). Most assessments of scientific literacy include analyses of scenarios that resemble scientific experiments and usually do not simulate scenarios that could be drawn from daily life. While previous research demonstrated that incorporating social context and background from students' everyday lives could improve learning (26), these factors most likely cannot explain why students in both the intervention and control courses showed improvement in our study. While students in the intervention group did have exercises that drew on everyday experiences (such as considering gluten intolerance), this context was lacking in the control course, although students still showed improvement. A difference that may account for the improvement is that the everyday-life scenario is more familiar to many students. In a previous study, the transfer of problem-solving ability was better when the problem included a more familiar introduction and context, rather than a less familiar context (27). One implication of this finding is that many assessments of scientific literacy may underestimate understanding by using scenarios that are unfamiliar to students.

An advantage of our intervention is that we improved the scientific literacy of experimental controls through a framework centered on the process of science and NOS. The work of Kyza on scientific literacy of early-secondary students is especially illuminating to our work (28). In a qualitative study of the scientific literacy of early-secondary students, Kyza examined how students interacted with a scaffolded process of inquiry using a software tool that especially emphasized developing and testing alternative hypotheses/explanations (28). While Kyza found that students showed learning gains after using the software, they had persistent epistemic problems, such as not easily developing alternative hypotheses or not understanding the

importance of developing alternative hypotheses (28). The work of Kyza suggests that interventions that lead to conceptual understanding of the epistemic basis of the process of science are necessary but not necessarily straightforward (28). We do find that students in our intervention course are more likely to describe confounding factors in their written explanations, suggesting that they are successfully considering alternative hypotheses in a broad sense (Fig. 4B).

Three factors strengthen the conclusions from our study. First, we examined student thinking with a two-tiered assessment that importantly included written answers in which students explained their thinking. The written answers were also scored blindly with respect to control and intervention groups by two independent scorers, limiting the effect of bias. We also found partial statistical support suggesting that the closed-ended and open-ended responses corroborated each other, giving us more confidence in our conclusions. Second, the same instructor taught both the control and intervention courses, and both had regular in-class active learning exercises integrated with the course content, so that the presence or absence of active learning or instructor variation is not expected to account for the differences observed. The instructor used active learning techniques in his courses and developed the curriculum for both the active learning (control) course and the subsequent intervention curriculum. By comparing two active learning strategies, this study contributes to the growing literature of research that goes beyond the first-generation studies on active learning (29), in which an active learning approach is compared to traditional instruction. Third, students in both the control and intervention courses were randomly assigned to those courses and had similar secondary school grades (high school GPAs) and university standardized examination (ACT test) scores, with similar distributions of students belonging to groups underrepresented in STEM fields (predominantly Latinx minorities and first-generation college students). Further, students in the two groups performed similarly in the precourse assessment. Thus, academic readiness, as judged by these metrics, probably does not explain the differential success in the control versus intervention groups.

While our study adds to the literature on scientific literacy and confirms some previous findings, we recognize limitations of our approach. Future studies should have larger sample sizes and should examine a wider diversity of undergraduate biology courses at more institutions. One factor that limited the size of our study was the analysis of written responses. However, because the closed-ended and open-ended responses corroborated each other, we may be able to use the closed-ended assessment with a larger number of students. An interesting future direction will be to focus on the scientific literacy of students who are not science majors.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, PDF file, 0.5 MB.

ACKNOWLEDGMENTS

We thank Andrew S. Johnson and Kayla C. Lewis for their insightful suggestions.

This work had no external financial support.

We declare no conflicts of interest.

REFERENCES

1. American Association for the Advancement of Science. 2011. Vision and change in undergraduate biology education: a call to action. final report. American Association for the Advancement of Science, Washington, DC. https://live-visionandchange.pantheonsite.io/wp-content/uploads/2011/03/Revised-Vision-and-Change-Final-Report.pdf.
2. Glass DJ. 2014. Experimental design for biologists, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
3. Giere RN. 1997. Understanding scientific reasoning, 4th ed. Harcourt Brace College Publishers, San Diego, CA.
4. Duschl R. 2008. Science education in three-part harmony: balancing conceptual, epistemic, and social learning goals. Rev Res Educ 32:268–291. https://doi.org/10.3102/0091732X07309371.
5. Dolan E, Grady J. 2010. Recognizing students' scientific reasoning: a tool for categorizing complexity of reasoning during teaching by inquiry. J Sci Teacher Educ 21:31–55. https://doi.org/10.1007/s10972-009-9154-7.
6. Shi J, Power J, Klymkowsky M. 2011. Revealing student thinking about experimental design and the roles of control experiments. Int J Scholarsh Teach Learn 5:Article 8.
7. Bennett KA. 2015. Using a discussion about scientific controversy to teach central concepts in experimental design. TEST 37:71–77. https://doi.org/10.1111/test.12071.
8. D'Costa AR, Schlueter MA. 2013. Scaffolded instruction improves student understanding of the scientific method & experimental design. Am Biol Teach 75:18–28. https://doi.org/10.1525/abt.2013.75.1.6.
9. Metz KE. 2004. Children's understanding of scientific inquiry: their conceptualization of uncertainty in investigations of their own design. Cogn Instr 22:219–290. https://doi.org/10.1207/s1532690xci2202_3.
10. Grunwald S, Hartman A. 2010. A case-based approach improves science students' experimental variable identification skills. J Coll Sci Teach 39:28–33.
11. Deane T, Nomme K, Jeffery E, Pollock C, Birol G. 2014. Development of the Biological Experimental Design Concept Inventory (BEDCI). CBE Life Sci Educ 13:540–551. https://doi.org/10.1187/cbe.13-11-0218.
12. Gormally C, Brickman P, Lutz M. 2012. Developing a Test of Scientific Literacy Skills (TOSLS): measuring undergraduates' evaluation of scientific information and arguments. CBE Life Sci Educ 11:364–377. https://doi.org/10.1187/cbe.12-03-0026.
13. 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. https://doi.org/10.1093/biosci/bit016.

14. 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. https://doi.org/10.1187/cbe.13-09-0192.
15. Killpack TL, Fulmer SM. 2018. Development of a tool to assess interrelated experimental design in introductory biology. J Microbiol Biol Educ 19:19.3.98. https://doi.org/10.1128/jmbe.v19i3.1627.
16. 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. https://doi.org/10.1187/cbe.10-01-0005.
17. Nelson CE. 1999. On the persistence of unicorns: the trade-off between content and critical thinking revisited, p 168–184. *In* Pescosolido BA, Aminzade R (ed), The social worlds of higher education: handbook for teaching in a new century. Pine Forge Press, Thousand Oaks, CA.
18. Finn KE, FitzPatrick K, Yan Z. 2017. Integrating lecture and laboratory in health sciences courses improves student satisfaction and performance. J Coll Sci Teach 47:66.
19. Luckie DB, Aubry JR, Marengo BJ, Rivkin AM, Foos LA, Maleszewski JJ. 2012. Less teaching, more learning: 10-yr study supports increasing student learning through less coverage and more inquiry. Adv Physiol Educ 36:325–335. https://doi.org/10.1152/advan.00017.2012.
20. Platt JR. 1964. Strong inference: certain systematic methods of scientific thinking may produce much more rapid progress than others. Science 146:347–353. https://doi.org/10.1126/science.146.3642.347.
21. Copi IM, Cohen C, McMahon K. 2016. Introduction to logic. Routledge, Abingdon, UK.
22. Hastie R, Dawes RM. 2009. Rational choice in an uncertain world: the psychology of judgment and decision making. Sage Publishing, Thousand Oaks, CA.
23. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. 2020. The proximal origin of SARS-CoV-2. Nat Med 26:450–452. https://doi.org/10.1038/s41591-020-0820-9.
24. Betini GS, Avgar T, Fryxell JM. 2017. Why are we not evaluating multiple competing hypotheses in ecology and evolution? R Soc Open Sci 4:160756. https://doi.org/10.1098/rsos.160756.
25. Sirum K, Humburg J. 2011. The Experimental Design Ability Test (EDAT). Bioscene J Coll Biol Teach 37:8–16.
26. Chamany K, Allen D, Tanner K. 2008. Making biology learning relevant to students: integrating people, history, and context into college biology teaching. CBE Life Sci Educ 7:267–278. https://doi.org/10.1187/cbe.08-06-0029.
27. Kole JA, Snyder HR, Brojde CL, Friend A. 2015. What's the problem? familiarity, working memory, and transfer in a problem-solving task. Am J Psychol 128:147–157. https://doi.org/10.5406/amerjpsyc.128.2.0147.
28. Kyza EA. 2009. Middle-school students' reasoning about alternative hypotheses in a scaffolded, software-based inquiry investigation. Cogn Instr 27:277–311. https://doi.org/10.1080/07370000903221718.
29. Freeman S, Eddy SL, McDonough M, Smith MK, Okoroafor N, Jordt H, Wenderoth MP. 2014. Active learning increases student performance in science, engineering, and mathematics. Proc Natl Acad Sci U S A 111:8410–8415. https://doi.org/10.1073/pnas.1319030111.