Title

Imagining, designing, and interpreting experiments: Using quantitative assessment to improve instruction in scientific reasoning

Permalink

https://escholarship.org/uc/item/6ww998mg

Journal

Biochemistry and Molecular Biology Education, 51(3)

ISSN

1470-8175

Authors

Coleman, Aaron B
Lorenzo, Kyla
McLamb, Flannery
et al.

Publication Date

2023-05-01

DOI

10.1002/bmb.21727

Supplemental Material

https://escholarship.org/uc/item/6ww998mg#supplemental

Received: 13 July 2022 | Revised: 5 March 2023 | Accepted: 20 March 2023

DOI: 10.1002/bmb.21727

ARTICLE

Biochemistry and Molecular Biology Education

WILEY

Imagining, designing, and interpreting experiments: Using quantitative assessment to improve instruction in scientific reasoning

Aaron B. Coleman[1] | Kyla Lorenzo[1] | Flannery McLamb[1] | Abhiraj Sanku[2] | Sahil Khan[3] | Goran Bozinovic[1]

[1]School of Biological Sciences, University of California, San Diego, California 92093, USA

[2]College of Osteopathic Medicine, Touro University, Vallejo, California 94592, USA

[3]School of Medicine, Albany Medical College, Albany, New York 12208, USA

Correspondence
Aaron B. Coleman, School of Biological Sciences, University of California, 9500 Gilman Drive, La Jolla, San Diego, California 92093-0355, USA.
Email: abcoleman@ucsd.edu

Abstract

Effectively teaching scientific reasoning requires an understanding of the challenges students face when learning these skills. We designed an assessment that measures undergraduate student abilities to form hypotheses, design experiments, and interpret data from experiments in cellular and molecular biology. The assessment uses intermediate-constraint free-response questions with a defined rubric to facilitate use with large classes, while identifying common reasoning errors that may prevent students from becoming proficient at designing and interpreting experiments. The assessment measured a statistically significant improvement in a senior-level biochemistry laboratory course, and a larger improvement between the biochemistry lab students and a separate cohort in a first-year introductory biology lab course. Two common errors were identified for forming hypotheses and using experimental controls. Students frequently constructed a hypothesis that was a restatement of the observation it was supposed to explain. They also often made comparisons to control conditions not included in an experiment. Both errors were most frequent among first-year students, and decreased in frequency as students completed the senior-level biochemistry lab. Further investigation of the absent controls error indicated that difficulties with reasoning about experimental controls may be widespread in undergraduate students. The assessment was a useful instrument for measuring improvement in scientific reasoning at different levels of instruction, and identified errors that can be targeted to improve instruction in the process of science.

KEYWORDS

assessment, controls, science-process skills, scientific reasoning

A college education is an invaluable asset that brings personal enrichment and better employment opportunities, and helps individuals to function in the democratic process as information-literate citizens. While subject knowledge is an important component of this, the skills and core competencies gained at university are equally if not more important in preparing students for the workforce and facilitating an intellectual engagement in society. This was the focus of the AAAS Vision and Change call for greater emphasis on research and the skills necessary to engage in the scientific process.[1] Many skills and core competencies we hope to foster in biology and biochemistry majors fall into the category of scientific reasoning, or science process skills.[2,3] While scientific reasoning captures a broad set of abilities, central to it is envisioning, designing, and interpreting experiments in a biological context. The emphasis on these skills is particularly important in laboratory courses, where students have the opportunity to conduct experiments and directly engage in the process of science.[4]

It is assumed that students enter college with some knowledge of how science is conducted, and that this represents a starting point to develop scientific reasoning in their biology coursework.[5] Hammer described these preexisting conceptions as resources from which students build more accurate and complete knowledge.[6] Undergraduate instruction has been criticized as student deficit-centered in not helping students to engage their prior knowledge.[7] However, as scientific reasoning encompasses a broad set of skills, it can be difficult for instructors to determine what the starting point should be for developing these abilities in their instruction.[2] Assessment is necessary to find the limits of preexisting conceptions and help the instructor gauge where to begin building from existing student understanding. Thus, quantitative assessment that can be used in classes with large enrollments is a critical aspect of teaching scientific reasoning at the university level. Assessment can also identify common roadblocks that students face in learning these skills. As these may reflect inaccuracies in or limitations to students' preexisting knowledge, they represent conceptions that are not yet fully developed, and hence we refer to them as incomplete conceptions.[6,8]

A range of instruments is needed for a comprehensive evaluation of students' reasoning abilities. There are published rubrics for the evaluation of these skills in students' writing for course assignments.[9,10] Here we focus on stand-alone assessments for measuring scientific reasoning outside of students' coursework. There is an abundance of assessments for measuring general scientific reasoning in primary and secondary school students.[11,12] Few assessments target university undergraduates, and fewer still apply these skills to content specific to the biological sciences. The Test of Integrated Science Process Skills (TIPS) by Dillashaw and Okay and the TIPS II test by Burns et al. were seminal instruments for the measurement of science process skills that were introduced in the 1980s.[13,14] While they were designed for middle and secondary school students and used questions that were not discipline-specific, they have been used for undergraduates in introductory biology courses.[15,16] Tobin and Capie published a similar instrument in 1982 that targeted a broader range of students up through college.[17] More recently, a handful of instruments specific for biology undergraduates have been developed. The Experimental Design Ability Test (EDAT) and an expanded version of the test (E-EDAT) use free-response, biology-related questions with a defined rubric.[18,19] The Biology Experimental Design Concept Inventory (BEDCI) uses biology-oriented, multiple-choice questions that do not require content knowledge.[20] While the EDAT and BEDCI assessments test scientific reasoning skills separate from any discipline-specific content knowledge, the MBDAT by Rybarczyk et al. measures these skills as applied to interpreting experiments in molecular biology.[21] The visual literacy for understanding how to read gel images and other types of discipline-specific data has been described as an important component of scientific reasoning in cell and molecular biology.[22] Likewise, the Neuron assessment by Dasgupta et al. incorporates representations in free-response essay questions on designing an experiment.[23]

Other assessments have focused on specific subsets of skills required for experimental design and interpretation. Shi et al. developed an instrument to measure how well students understand experimental controls, and we have previously published an assessment for the ability to use the concepts of correlation, necessity, and sufficiency in interpreting cell and molecular biology data.[24,25] Two other assessments measure different but related science process and literacy skill sets; the BioSQuaRE assessment measures quantitative reasoning ability in a biology context and includes questions on reading common types of quantitative data representations, such as heat maps, while the TOSLS assessment focuses on information literacy but overlaps with the others in measuring the ability to interpret graphs and other data sets.[26,27]

No one assessment can address the full range of scientific reasoning skills we aim to improve in our undergraduates. Different instruments are required to measure the understanding of hypothesis formation and experimental design in introductory biology classes versus the application of these skills in advanced classes where the integration of content knowledge is required. Both are necessary to ensure that students' progress from novice to expert-like thinking. The ability to interpret and utilize data in scientific reasoning may be particularly important in this progression. In comparing two versions of an assessment to measure manipulation of experimental variables in physics that utilized matched questions that either included or left out data interpretation, including data increased the difficulty.[28] The authors speculated this came from the need for students to use higher-order cognitive skills to process the questions that included data. The upper-level Bloom's taxonomy skills of analysis, synthesis, and evaluation may be required to bridge simply understanding definitions (e.g., what controls are for), to using reasoning skills in working with real-life, discipline-specific data.[29] Problems that require the integration of multiple cognitive components more accurately reflect what experts do in conducting research.

Beyond verifying improvement in scientific reasoning, assessment should identify specific difficulties students face in learning these skills. By identifying common reasoning errors, instructors can adjust their teaching to help students overcome them. Some published studies have made progress in this regard. Shi et al. found that students often do not understand the role of controls in the interpretation of experimental results.[24] We found that students often incorrectly interpret a correlation as causation in the form of sufficiency when analyzing discipline-specific data.[25]

The format of the assessment questions impacts how much information they can provide about particular reasoning actions that pose difficulty. Instruments that use multiple-choice questions are easier to administer to large groups of students but do not easily measure higher-order Bloom's learning.[30] Open-ended free-response (essay) questions provide a deeper understanding of student thinking and are more likely to identify specific areas of difficulty, but due to the time required to score the questions they are generally not feasible for assessing high-enrollment classes.[31] Intermediate-constraint free-response questions, which limit the length and form in which students can answer while still requiring higher-order cognitive actions, offer a reasonable compromise.[32,33]

We designed a scientific reasoning skills assessment to measure students' abilities to form hypotheses to explain a given set of observations and to design and interpret experiments to test a given hypothesis. These are skills we focus on in our courses, however they are broadly applicable, and the assessment was created for general biology undergraduate use. Our intent was to produce an instrument with questions of increasing difficulty to quantify a range of skill levels as students' progress to more expert-like thinking. This is distinct from the goal of a concept inventory, where the purpose is to test whether students have achieved a defined set of learning outcomes.[34] We also wanted to identify common reasoning errors hindering students from becoming proficient at these skills. We incorporated both multiple-choice and intermediate-constraint free-response questions into the instrument to better understand students' reasoning while conceiving and designing experiments. The assessment revealed statistically significant improvement in students' scientific reasoning from the beginning to end of an upper-division biochemistry lab course. It also demonstrated stronger reasoning skills in the biochemistry lab students when compared to first-year students in an introductory biology lab class. The students' responses provided insight into aspects of forming hypotheses and designing experiments that were challenging. We found that students frequently made comparisons to control conditions not included in an experiment, and implemented a shorter secondary assessment in a separate cohort to further investigate this reasoning error. We provide evidence that incomplete conceptions about experimental controls may be widespread and complex.

1 | METHODS

1.1 | Study overview

The primary assessment was designed and validated using undergraduate classes at a large, high-enrollment research university in Southern California. It was implemented in an upper-division biochemistry laboratory course (**UD-Lab**) and then a lower-division, introductory biology lab course (**Intro-Lab)**. Using the data generated by the primary assessment, we employed a grounded theory approach to further investigate common reasoning errors that were detected.[35,36] Frequent mistakes that were identified on the free-response questions were coded and quantified. One common error was targeted for further study to better determine its prevalence in biology undergraduates and elucidate the specific cognitive actions involved. A shorter, secondary assessment that focused on this error was developed and administered to a broader cohort in a biochemistry lecture course (**Biochem-Lec**).

1.2 | Courses

UD-Lab: The upper-division biochemistry lab course is a techniques-based class that focuses on the purification and analysis of proteins. The curriculum emphasizes quantitative and analytical reasoning skills. Students were given the primary scientific reasoning skills assessment in seven biochemistry lab classes from fall 2017 to winter 2019, taught by three different instructors (Table S1). The UD-Lab classes have either 3 or 6 lab sections that enroll up to 24 students each, and 58%–96% of the students in each class completed both the pre and post-class assessment (mean participation was 81%). **Intro-Lab:** The introductory biology lab course is taken by biology majors in their first-year and teaches foundational laboratory skills and biological concepts. This is a high-enrollment course, and the assessment was administered to a single class of 190 students in fall quarter 2019 with an 85% participation rate. **Biochem-Lec:** The secondary assessment was implemented in a high-enrollment metabolic biochemistry lecture course to capture a larger number of students more quickly. This was a single class of 518 students in spring quarter of 2021, and the participation rate in the secondary assessment was 56%. This course has no laboratory component, and this particular class was taught remotely due to the COVID-19 pandemic. More detailed descriptions of each course and a partial breakdown of their demographics is provided in the supplemental information.

1.3 | Development of the primary assessment

The genesis of this instrument was the authors' observations of our students' scientific reasoning in the UD-Lab class. To develop the assessment, we first compiled a list of broad skills that we wanted our students to master; for example, design an experiment with the appropriate controls (Table 1). Specific cognitive actions necessary for each skill were then filled in, with an emphasis on where we observed frequent mistakes. These skills and their associated cognitive actions are widely accepted to be central components of scientific reasoning.[1,2,10] We wanted the instrument to capture a range of improvement in these skills as students progressed from more-novice to more-expert like ability and not merely whether they could do the reasoning tasks we gave them in the class, and we wanted it to inform us about the difficulties that hindered this progression. To accomplish this, we estimated the starting point at which students would begin to build on existing skills and conceptions upon entering the UD-Lab.[6] Question sets were structured to first assess these preexisting skills and then increasingly difficult reasoning tasks that we hoped they would develop in the UD-Lab and beyond. To determine common reasoning errors, we used intermediate-constraint free-response questions that guided the direction and length of the response. Rubrics were designed for these questions that awarded points independently for various correct elements in the answers, which allowed us to measure a range of ability with fewer questions. Other goals guiding the design of the instrument were that it should: (a) require students to use the

TABLE 1 Design of the primary scientific reasoning skills assessment.

Skill	Question number	Point value[a]	Format	Content knowledge needed	Predicted cognitive actions tested
Form hypothesis	1	2	free response	None	Identify dependent and independent variables.
	2	2	free response	None	Identify facts relevant to an observation; identify patterns of causation.
Interpret data	3a	1	multiple choice	Western blotting	Visual reasoning/interpret representation; understand time course.
	3b	1	multiple choice	Western blotting	Visual reasoning/interpret representation; distinguish experiment showing causation vs. correlation.
	4	2	select amino acids in given sequence	Protein amino acid sequence	Visual reasoning/interpret representation; apply concept of competition (explained in question) in biochemical interactions.
Design experiment	5	4	free response	None	Minimize variables between conditions; utilize controls to repeat preestablished observations; predict result for conditions based on hypothesis.

[a]The point weighting for each question out of 12 points total for the assessment.

higher-order cognitive skills needed for scientific reasoning in an actual research context; (b) require minimal content knowledge; (c) be completable in less than 1 h; and (d) allow for scoring that was rapid enough to make it useable in large classes.

An initial version of the assessment was given to a UD-Lab class, and student responses guided modification of the questions and rubric. The final version of the assessment contained five questions that were a combination of multiple choice and free response (Table 1). Points were weighted to fix the question sets for hypothesis formation, data interpretation, and experimental design at 4 points each. The full assessment with rubric is provided in the supplemental information.

Questions 1 and 2 addressed the ability to form testable hypotheses. The design of these questions came from observations that the UD-Lab students generally understood a hypothesis proposed potential new knowledge, but were unclear on what it defined. They would sometimes state a hypothesis as the experiment they were going to conduct. Question 1 was adapted from the TIPS tests,[13,14] as used by Dirks and Cunningham.[15] The original multiple-choice format was changed to free response, and this question measured the ability to assign dependent and independent variables in a hypothesis statement. Question 2 examined how students sort relevant information when forming a hypothesis to explain an observation. It was intended to get beyond simply assigning variables and get at the higher-order cognitive actions (create/synthesize Bloom's level) scientists perform when processing information, particularly in the ability to establish patterns. The question stems from the vitamin D-folate hypothesis, which predicts skin pigmentation balances UV-light mediated vitamin D synthesis versus folate breakdown.[37] It lists a series of facts and observations regarding vitamins and vitamin deficiencies and then prompts students to write a hypothesis to explain one of the observations. The rubric assigns points for establishing and connecting relevant facts in forming the hypothesis.

From our experience in the UD-Lab, most students could perform the individual reasoning tasks necessary for interpreting an experiment; for example, reading Western blot band intensities for the abundance of a protein, or understanding how an experimental reagent would affect the protein. However, they frequently had trouble combining multiple reasoning tasks, and struggled when asked to combine their knowledge of Western blotting and the experimental reagent to identify a predicted banding pattern. Therefore, questions 3 and 4 measured students' abilities to interpret data and derive the result of an experiment where they had to integrate knowledge of the technique with understanding of the experimental design. Some degree of preexisting content knowledge was required for these questions, however interpreting discipline-specific data representations is an integral part of scientific reasoning.[22,28] It was deemed most students would have this content knowledge upon entering the UD-Lab class, and this knowledge would be expected of all biology undergraduates upon graduation. More constrained question formats were used, as it was difficult to devise free-response questions that could be rapidly scored without making the questions too leading. Question 4 had students interpret a published experiment that used competition to investigate integrin-binding.[38] The concept of competition was explained in the question.

To examine how students design experiments to test a given hypothesis, we used a single question (question 5) that had multiple, independently-scoreable components. In the UD-Lab course, students get to design an experiment, and it is common for them to omit necessary control conditions and sufficient experimental replicates. They also frequently struggle with predicting the result of their experimental design, given their hypothesis is true. In question 5 we tried to cover as many of these cognitive actions as possible. It provided background information and a hypothesis about the egg laying preference of bean beetles,[39] and then asked students to design an experiment to test the hypothesis. It tested the ability to minimize unwanted variables and establish control conditions by prompting them to define what they would add to four experimental dishes, and tested the ability to address the hypothesis by having them make a prediction about the number of eggs laid in each dish relative to the other dishes.

1.4 | Implementation of the primary assessment

For the UD-Lab, the assessment was administered on the first day of lab (pretest) and during the last week of the 10-week quarter (posttest). For the Intro-Lab, the assessment was given once during the first lab session. On each occasion, the students were given 45 min to complete the pencil and paper assessment. Students were incentivized to participate in the study either by earning extra credit that was up to 0.4% of the total class points for both the pretest and posttest, based on how many questions they answered correctly, or they were told that taking the assessment would help them prepare for the quizzes and exams in the class. The study was carried out with institutional review board approval from the UCSD Human Research Protections Program (project #171306XX).

The assessments were scored by the lead author (Intro-Lab and two of the 7 UD-Lab classes; Table S1) or by one of two research students who scored two and three of the UD-Lab classes, respectively. The rubric, short length, and focus of the free-response question answers facilitated consistent scoring.

1.5 | Validation of the primary assessment

Two forms of interviews were conducted with students after they had completed the assessment. To validate that the questions were testing the expected cognitive actions, three students from a winter 2021 UD-Lab that was not part of the study group participated in in-depth interviews. These students were volunteers and their participation was not incentivized. They completed the assessment electronically and were then interviewed over Zoom, during which they were asked to walk the interviewer through their thought process for answering the questions. Question 3 was omitted from the in-depth interviews for brevity. For question 2, interviewees were additionally asked to explain why they included or excluded each rubric element. The interviews were not recorded and the interviewer typed notes on student responses during the interview.

To gain a broader question validation, 16 students participated in opinion-survey interviews where their answers were captured on a four-point Likert scale. For each question (the two sub-questions for question 3 were surveyed separately), the students were first asked if the question was clear (very clear to not at all clear) and their level of confidence in their answer (very confident to not at all confident). To address the solvability of the questions, the interviewer then explained the answer and scoring followed by asking the likelihood that a student who was very good at that skill would get all the points for the question. The Likert scale was: very likely (90% or more of the time), somewhat likely (more than 50% of the time), not very likely (less than 50% of the time), and not at all likely (less than 10% of the time). For questions 2 and 5 that had multiple scoring elements defined in the rubric, the different elements were addressed by separate survey questions. In Figure 1, responses for the multiple rubric elements have been averaged to get a single value for solvability. Six students from the winter quarter 2019 UD-Lab class that was part of the study were interviewed immediately after taking the post-class assessment. To indicate how applicable the assessment is for students outside the UD-Lab course, 10 students from a BiochemLec class (not part of the study) were interviewed in spring quarter 2020.

1.6 | Development and implementation of the secondary assessment

A secondary assessment was developed to further investigate a common reasoning error about how experimental controls are used. This assessment had only two questions, question 5 from the primary assessment and one new question. For simplicity, the question numbering established on the primary assessment was carried over to the secondary assessment. Thus, question 6 is the new question and question 5 is the same as the primary assessment, although on the version taken by students these were presented as questions 1 and 2, respectively. Three different versions of question 5 were used on the secondary assessment to better understand how the question structure affected student answers. Each version was given to a different group of students. One version was identical to the primary assessment question 5, a second version predefined the experimental samples (contents of the dishes), and a third added another potential control sample (a third bean of known egg-laying preference). Question 6 was tailored to determine if the experimental controls reasoning error also applied to understanding, what controls are necessary to make a valid interpretation of experimental data. It used a multiple-choice format to have students identify the correctly controlled experiment.

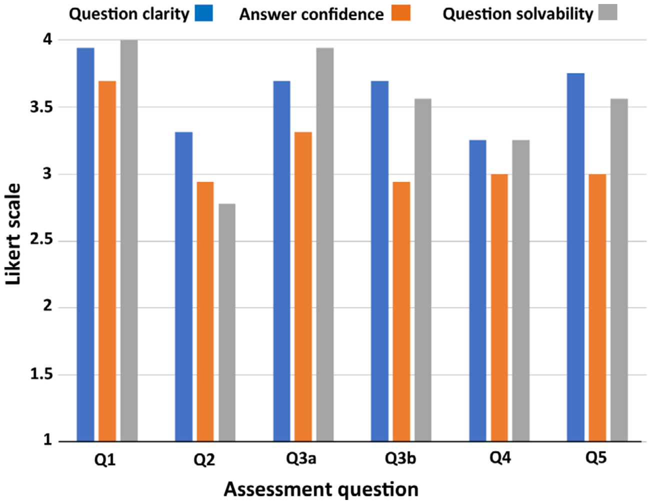

FIGURE 1 Assessment validation interviews. Student interviews were conducted where the interviewee was guided through a four-point Likert survey to measure their opinion of the clarity of each question (*very clear* = 4 to *not at all clear* = 1; blue bars), their confidence in their answer (*very confident* = 4 to *not at all confident* = 1; orange bars), and how likely it would be for a student who was very good at that reasoning skill to get full points for that question (*Very likely [90% or more of the time]* = 4 to *Not at all likely [less than 10% of the time]* = 1; gray bars). Bars represent the averaged responses from 16 interviews. For assessment questions scored for multiple rubric elements, the students were asked about each rubric element independently (two for question 2 and four for question 5). For these questions, the question solvability bar represents student responses averaged across the multiple rubric elements.

The 2-question secondary assessment was implemented in a spring 2021 Biochem-Lec class. It was administered asynchronously using Qualtrics, and students were given a 10-day window to take it with no time limit to complete the questions once they began the assessment. The three different versions were given to different discussion sections in the class by providing alternate links. This assessment was also incentivized by offering up to 0.4% of the class total points in extra credit.

1.7 | Data management and statistical analysis

To determine how the assessment would measure increasing differences in scientific reasoning skill level, pairwise tests were performed between UD-Lab pretest, UD-Lab posttest, and Intro-Lab scores. For the UD-Lab, only matched pre- and posttest assessments were included for paired analysis, while unpaired tests were performed between the Intro-Lab and the UD-Lab pretest, and between the Intro-Lab and the UD-Lab posttest. Due to non-normal distributions as determined by Shapiro–Wilk tests, assessments were analyzed by total and by individual question scores via two-tailed Wilcoxon signed-rank tests with Holm's correction. Effect sizes were calculated as Cohen's *d*.

Unscoreable and blank assessment responses were treated as a score of 0. To determine how unscoreable and blank responses to the highly weighted question 5-affected analysis, paired Wilcoxon tests of total pre- versus posttest scores were repeated with unscoreable and blank question 5 responses dropped. The total for those students were scored out of 8, and then renormalized to 12 to remove any penalty for question 5. All Shapiro–Wilk, Wilcoxon, and effect size calculations were performed in R version 4.1.2.

To examine score similarities for questions from paired pre and posttests (UD-Lab only), version 1.2.1 of the heatmaply package in R was used to hierarchically cluster pre and posttest questions together by score to produce a combined heatmap and dendrogram.

To quantify the frequency of specific reasoning errors, a subset of the assessments from the UD-Lab were reevaluated, along with all the assessments from Intro-Lab. For question 2, the frequencies were determined from 86 matched pre and posttest assessments from the spring 2019 UD-Lab class. For question 5, the frequencies were determined from 367 matched pre and posttest assessments from the UD-Lab, excluding winter 2018 and fall 2018. To test for association between committing the absent control error on Q5 and answering Q6 incorrectly, we ran chi-square tests in R. A nominal threshold of *p*-value < 0.05 was set for significance on all tests.

2 | RESULTS

We ran the primary scientific reasoning skills assessment in seven different classes of the UD-Lab course, taught by three instructors, from fall 2017 to spring 2019 (Table 2 and Table S1). Taking all 7 classes in aggregate, 470 students completed the assessment out of 591 enrolled. The pretest average was 51.4% and the posttest average was 55.1% with an effect size of 0.24 (adjusted $p < 0.001$). In scoring the assessments, we noticed that a small but consistent number of students in each class did not follow the directions for question 5 and answered in a way that could not be scored with the rubric (for example, placing more than one type of bean in each dish). A smaller number of students completed the other assessment questions but provided no answer for question 5. Together, these occurred on 7% of the pretests and 5% of the posttests. In either instance the question was scored as zero out of 4 points. Due to the large weighting of question 5 relative to the other questions, we wanted to ensure these assessments were not disproportionately affecting the results. For assessments where question 5 was either not scoreable or unanswered, question 5 was dropped and that assessment was given a score out of 8 points, which was then normalized to 12 points. This gave pretest and posttest averages of 52.7% and 56.2%, respectively, with an effect size of 0.24 (adjusted $p < 0.001$). As this did not affect the interpretation of the data, further analysis was done with unscoreable question 5 answers marked as zero.

Separating the assessment by score quartile showed a trend toward improvement from pretest to posttest (Figure 2a). On the pretest, 58% of the students scored in the top two quartiles and this improved to 68% of the students on the posttest. Pretest to posttest comparison for individual students showed that not all students improved (Figure 2b); 12% of the students showed no change, 53% improved, and 36% did worse. However, 16% of the students improved their posttest score by 20% or more, while only 6% of students had posttest scores that decreased by this amount, suggesting that the change in aggregate averages reflects a real improvement. This change was driven primarily by improvement on questions 4 and 5 (Table 2 and Figure 3). Individual student scores for each question were analyzed to determine if they followed an overall pretest to posttest pattern. The score similarities for each question, pretest and posttest, were quantified for each student and clustered hierarchically by score. The analysis showed that the pretest and posttest scores for each question clustered together, as expected. Question 1 had the highest average and showed the greatest similarity in students' pre and posttest scores, while question 4, for which the average increased the

TABLE 2 Assessment of scientific reasoning skills in undergraduate biology students.

	Upper-div. biochemistry lab course (UD-Lab)		Freshman intro-biology lab course (Intro-Lab)	
	470 students			
Question	**Pretest**	**Posttest**	**162 students**	**Statistical significance**
1	94.7%	96.8%	97.7%	NS
2	21.3%	22.2%	13.6%	UDPre-Intro**; UDPost-Intro*
3a	93.6%	93.8%	85.2%	UDPre-Intro***; UDPost-Intro***
3b	44.0%	44.5%	31.5%	UDPre-Intro**; UDPost-Intro**
4	42.6%	51.9%	19.8%	UDPre-UDPost***; UDPre-Intro***; UDPost-Intro***
5	40.6%	45.4%	38.8%	UDPre-UDPost***; UDPost-Intro***
Total	51.4%	55.1%	44.5%	UDPre-UDPost***; UDPre-Intro***; UDPost-Intro***

Note: The aggregate assessment results are shown for seven biochemistry lab classes (UD-Lab) from fall 2017 to spring 2019, and for a single freshman introductory biology lab (Intro-Lab) in fall 2019. For each round of the assessment, the average score is given as the percentage of the points possible for that question or for the assessment total. Statistical significance was determined by two-tailed Wilcoxon signed-rank tests with Holm's correction for non-normal distribution of the data. Significance of pairwise comparisons of the UD-Lab pretest (UDPre), the UD-Lab posttest (UDPost), or the Intro-Lab (Intro) are indicated as * adjusted $p < 0.05$, ** adjusted $p < 0.01$, *** adjusted $p < 0.001$, or not significant (NS) if none of the comparisons were significant.

most, showed the least similarity. The clustering pattern for individual scores generally followed the same pattern as the average scores, suggesting the pretest to posttest differences for questions 4 and 5 represent actual changes in student performance.

Since the small pretest to posttest improvement might be expected over the course of a single class, we wanted to validate that the assessment could capture broader differences in skill levels. A single round of the assessment was given to the students in a freshman-level introductory biology lab class (Intro-Lab) in fall 2019 (Table 2). Out of 190 students enrolled, 162 completed the assessment. The average score for this cohort was 44.5%. Comparing the assessment performance of the Intro-Lab students to the UD-Lab students showed significant differences in the average for both the UD-Lab pretest (adjusted $p < 0.001$), with an effect size of 0.51, and the UD-Lab posttest (adjusted $p < 0.001$), with an effect size of 0.77. Blank or unscoreable question 5 answers occurred on 6% of the Intro-Lab assessments, and eliminating these gave similar, significant differences with the UD-Lab pretest and posttest (adjusted $p < 0.001$). The distribution of assessment scores showed fewer students in the top two quartiles relative to the UD-Lab and more students in the bottom two quartiles (Figure 2a). The Intro-Lab students did not perform as well as the UD-Lab students on four of the five individual assessment questions (Table 2).

An examination of each assessment question independently provided a better picture of how the questions were functioning and how well students were learning particular skills. Questions 1 and 2 addressed the ability to form a testable hypothesis to explain a given observation. While question 1 tested basic understanding of dependent and independent variables, question 2 tested higher-order cognitive skills needed to parse complex sets of information and synthesize a model that explains an observation. Most students across all three rounds of the assessment in both classes gave a fully correct answer and received all the points for question 1 (Table 2). However, for question 2 most students received zero or a fraction of the possible points. Points were received for at least one rubric element by 40% of the students in the Intro-Lab, and by 48% on the UD-Lab pretest and 46% on the UD-Lab posttest. Question 2 performance did not change significantly from pretest to posttest in the UD-Lab, but did show a significant difference between the Intro-Lab cohort and both the pretest (adjusted $p < 0.01$) and the posttest (adjusted $p < 0.05$) for the UD-Lab.

In-depth validation interviews with three students suggested they were engaging in the expected cognitive actions for both questions 1 and 2. For question 2, all three students appeared to weigh the different facts presented in the question and tried to make connections between them. They differed, however, in how they interpreted the facts and their confidence in their answer. Two students whose answers were scored as zero out of 2 points indicated they had low confidence in their

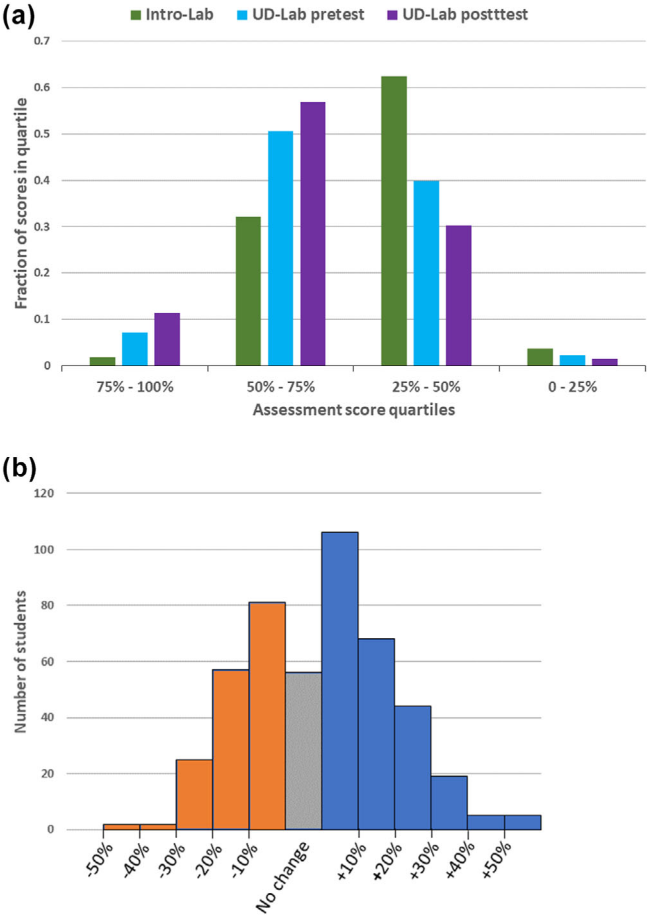

FIGURE 2 Analysis of the primary assessment score distributions. (a). The assessment total scores are shown by quartile for the Intro-Lab (162 students) and the UD-Lab (470 students) pretest and posttest. Bars indicate the number of assessment total scores that fall into each quartile, as a fraction of the total number of assessments for each group. (b). The pretest to posttest differences in the total assessment scores for individual students in the UD-Lab are shown as a histogram. Bars indicate the number of students with each change, with no change shown in gray, a positive change shown in blue, and a negative change shown in orange.

answer, and the third student who received 2 out of 2 points indicated they were confident in their answer.

Opinion-survey validation interviews were conducted with an additional 16 students (Figure 1). Students found question 2 somewhat to very clear, in what it was asking (3.3 average response). The rubric for question 2 defines multiple elements that would come together to form the best hypothesis, and the interview survey combines them into two sets of relevant information (greater sunlight UV absorption and UV light breaks down folate) that are addressed in separate survey items. When the answer and scoring rubric were explained, students found it relatively likely that a student who was good at this skill would include greater UV absorption (3.1 average response) but less likely that UV breaks down folate would be included (2.4 average response), giving a combined average of 2.8. To further investigate the validity of question 2, we looked at how many students gave the full correct answer and received the maximum score. Three students (1.9% of 162) in the Intro-Lab, 28 students on the UD-Lab pretest (6.0% of 470), and 28 students on the UD-Lab posttest (6.0% of 470) received the maximum score. If question 2 was working as expected, we predicted there would be a positive association between the total score on the assessment and the question 2 score. The posttest assessments from the UD-Lab classes showed a Pearson's r value of 0.38 for this comparison, indicating a moderate positive correlation between performance on question 2 and the assessment as a whole.

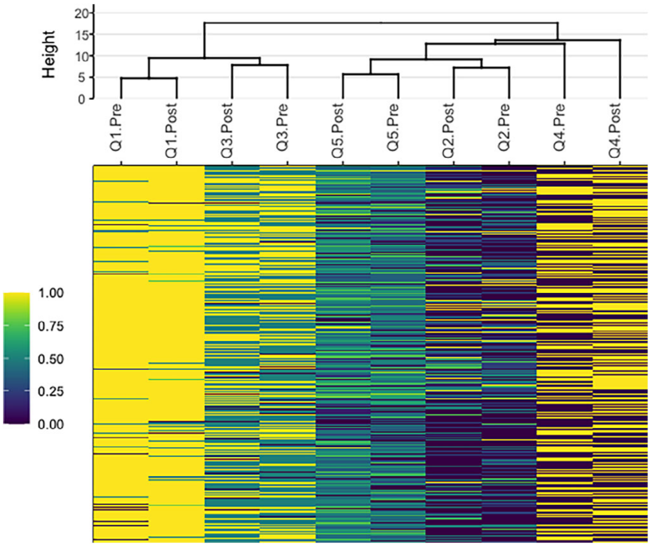

FIGURE 3 Comparison of individual student scores by question. Individual student scores (vertical axis) were sorted for each question (horizontal axis) and clustered by score similarity and presented as a heat map. Similarities are clustered hierarchically, with greater similarity on the left and lower similarity on the right. The color key indicates the score for each question as a fraction of the points possible. The two sub-questions for question 3 were analyzed as a single question. The dendrogram above the question columns indicates the clustering pattern, with greater branch height indicating greater score dissimilarity.

Question 2 provides the observation "Public health officials have observed that light-skinned individuals living in the tropics have a greater risk for neural tube defects during pregnancy," and then prompts students to explain this relationship. Rather than proposing a hypothesis to explain the observation, some students answered by simply restating the observation (Table 3). The first-year students in Intro-Lab committed this error 38.3% ($n = 162$) of the time, and it occurred at lower frequencies with senior-level students ($n = 86$) on the UD-

TABLE 3 Frequencies of identified reasoning errors.

Question	Reasoning error identified	Upper-div. biochemistry lab course		Freshman intro-biology lab course (sample size)
		Pretest	Posttest	
2[a]	Hypothesis as restatement of the observation	29.1% ($n = 86$)	23.3% ($n = 86$)	38.3% ($n = 162$)
5[b]	Making comparisons between samples and absent controls	19.3% ($n = 367$)	13.9% ($n = 367$)	22.8% ($n = 162$)

Note: The error frequency and sample size (n) are given for each round of the assessment.

[a]Frequencies were determined from the spring 2019 UD-Lab class and from the fall 2019 Intro-Lab class.

[b]Frequencies were determined from 5 UD-Lab classes and from the fall 2019 Intro-Lab class.

Lab pretest (29.1%) and UD-Lab posttest (23.3%). One of the three students who participated in an in-depth validation interview gave this incorrect answer for question 2, writing "A lighter skin color is associated with an increased risk for neural tube defects." When asked how confident they were in their answer, the student indicated they were not confident because they did not understand the relationship between vitamin B (folate) and skin coloration. The prevalence of this incorrect answer could indicate a common incomplete conception about how a hypothesis should form the basis for an experiment or study to provide new information rather than a formal restatement of what is known.

Questions 3 and 4 examined students' abilities to interpret data representations and determine the result of an experiment (Table 2). Question 3a gave Western blot results for a time-course experiment and asked the order in which expression of two proteins is turned on. All three rounds of the assessment showed relatively high performance for this question. There was no difference between the pretest and posttest for the UD-Lab (93.6% and 93.8%, respectively). The Intro-Lab was 85.2%, and this was significantly different from both the UD-Lab pretest (adjusted $p < 0.001$) and posttest (adjusted $p < 0.001$). Question 3b had students interpret another Western blot for an experiment designed to determine if an intermediate factor causes a biological effect, and tests their understanding of causation versus correlation. This question proved more challenging, with averages of 44.0% and 44.5% for the pretest and posttest in the UD-Lab, respectively, and 31.5% in the Intro-Lab (adjusted $p < 0.01$ for Intro-Lab vs. both UD-Lab pretest and posttest). Question 4 had students interpret a competition experiment to determine the amino acid sequence that forms a protein binding site. The concept of competition disrupting biochemical binding interactions is explained in the question. Answering the question requires interpreting an x, y-plot showing the effect of increasing competitor concentration and requires some content knowledge of protein amino acid sequences. This question showed the largest changes in performance across the different rounds of the assessment. The average on the UD-Lab pretest was 42.6% and improved to 51.9% on the posttest (adjusted $p < 0.001$). The Intro-Lab average was only 19.8% and this was significantly less than both the UD-Lab pretest and posttest (adjusted $p < 0.001$ for both comparisons). Student responses in the in-depth validation interviews suggested question 4 was working as expected. In the survey interviews, students found the clarity and solvability of questions 3a and 3b to be relatively high, with average Likert responses greater than 3.5 for both (Figure 1). They rated question 4 lower on these items, with an average response of 3.3 for both clarity and solvability.

Question 5 measured reasoning skills for designing experiments to test a given hypothesis. It provides a brief background and hypothesis, and directs students to define four experimental conditions and to make a prediction about each condition given that the hypothesis is true. It is scored for multiple rubric elements and allows for a broader point distribution (out of 4 points) than the other questions. The UD-Lab pretest average was 40.6% which significantly increased to a posttest average of 45.4%, (adjusted $p < 0.001$). Examining pretest to posttest changes in individual scores showed that 39% of the students improved their score, 36% showed no change, and 26% did worse (Figure S1 panel B). The Intro-Lab average was 38.8%, which was not significantly different from the UD-Lab pretest, but was significantly different from the posttest (adjusted $p < 0.001$). When unscoreable or blank answers were omitted from the analysis, this gave averages of 44.0%, 47.9%, and 41.1% for the UD-Lab pretest and posttest and the Intro-Lab, respectively, and it did not affect the statistical significance of the changes between rounds of the assessment. The overall performance on the question showed a trend toward improvement across all three rounds of the assessment (Figure S1 panel A). Although the average scores were not significantly different, the percentage of students scoring in the highest quartile was 5.6% for the Intro-Lab and 10.2% for the UD-Lab pretest.

In the in-depth validation interviews for question 5, the students thought mostly about which beans to select for their experimental conditions and what they could predict from these conditions. When asked if there were any aspects of the question they might not have accounted for in their answer, one of the three students was not confident about which variables could be changed in the experimental conditions. In the survey interviews, the average Likert responses were greater than 3.5 for both question clarity and the composite ranking of question solvability (Figure 1). The response differed, however, for one of the four rubric elements making up the composite question solvability ranking. The rubric element for quantifying sample contents, included to assess the ability to minimize unwanted variables, received an average Likert response of 3.0. This survey response indicated they thought it was only somewhat likely that a student who was very good at designing experiments would quantify the sample contents in their answer. The other three rubric elements for solvability were all ranked 3.5 or higher. This tracked with poor performance on the quantifying sample contents rubric item in the assessment. The question 5 prompt to define the experimental conditions states: "List exactly what you would place into each of the four dishes in the spaces provided below, as if you were really going to perform the experiment." Most students did not quantify the contents of the experimental dishes. The rubric allocated 1 point (out of 4 total) for this element. It is not clear, if this common omission represented a lack of understanding about minimizing unwanted variables, or if it simply did not occur to students to do this in the context of answering the question. In the survey-interviews, one student commented that they thought quantity was implied.

Another common mistake in students' answers for question 5 more clearly indicated an error in reasoning. The question provides information that allows students to select control conditions when they define their experimental samples (bean varieties with known high and low egg-laying preference for bean beetles). When indicating their prediction for the result for each sample, students would frequently make comparisons between the sample and a control sample that they did not include. For example, one student chose lima, pinto, navy beans, and lentils for their four samples, but then made a prediction for each about the number of eggs laid relative to kidney and mung beans, which were described in the question as having high and low egg-laying preference. Students making this error appeared to assume that, although high and low egg-laying preference was the only information given, this information provided a fixed point of reference to which any future sample could be compared without including the controls in the experiment. This error was quantified on assessments from 5 of the 7 UD-Lab classes and for the Intro-Lab class (Table 3). The error occurred on 22.8% of the assessments in the Intro-Lab ($n = 162$), 19.3% of the UD-Lab pretests ($n = 367$), and 13.9% of the UD-Lab posttests ($n = 367$).

To further investigate the prevalence of this error and to better understand student thinking about how controls are used, we developed a secondary assessment instrument containing only two questions. This instrument included question 5 from the primary assessment and one new question (question 6). Two modified versions of question 5 were created, and these plus the original version were given to different groups of students.

Question 5 placed constraints on the number of experimental samples that could be run. We hypothesized that the frequency at which students made erroneous comparisons to absent controls would increase if the constraints on sample selection were further increased. Two versions of the question were given to students in a Biochem-Lec class in spring 2021; on one version the samples were preselected and did not include the beans with known egg-laying preference (high constraint) and the other version was the original question where the students selected the samples (moderate constraint). 96 students who received the high-constraint version made comparisons to absent controls at a frequency of 27%, and 79 students who received the original, moderate-constraint version made comparisons to absent controls at a frequency of 14% (Table S2). Additionally, 98 students received a third version that added another potential control that could be selected for the experimental samples (another bean of known egg-laying preference). This group made comparisons to absent controls at a frequency of 15%.

Secondary assessment question 6 tested the ability to distinguish a correctly controlled experiment from an incorrectly controlled experiment. With the premise of testing if an inhibitor would block cell growth stimulation by epidermal growth factor (EGF), each of the four multiple choice answers showed the results of an experiment where four conditions were run. The correct answer (b) included a control where nothing was added to the cells, establishing that growth stimulation occurred when EGF was added, but showed no inhibition of this growth stimulation with the inhibitor. An incorrect answer was devised to test if students would erroneously assume that preestablished information created a fixed point of reference for measuring growth stimulation. This answer (d) lacked a control where nothing was added and therefore did not establish the EGF was stimulating growth, but then showed less growth when the inhibitor was added with EGF. The two other incorrect answers both included a control where nothing was added, but lacked a condition where EGF and the inhibitor were

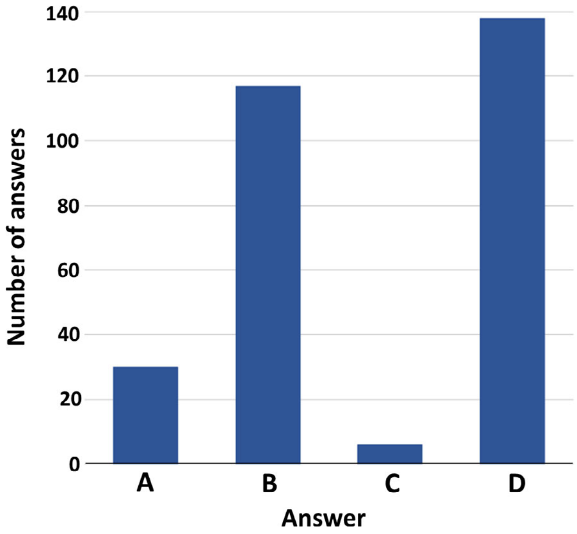

FIGURE 4 Students make frequent invalid interpretations from uncontrolled experiments. Three hundred and nineteen students in a single upper-division biochemistry lecture class (Biochem-Lec) were given a new assessment question that targeted their ability to identify a correctly controlled experiment. Answer B is the correct answer; the experiment has the necessary controls and shows no effect (indicates the null hypothesis for the experiment is true). Answer D is an incorrect answer; the experiment is missing a control needed to make a valid interpretation and shows an effect (indicates the hypothesis for the experiment is true). Bars represent the number of students giving each answer.

added together. Out of all 291 students who took the secondary assessment, 47.4% selected incorrect answer d that lacked the necessary control, and 40.2% selected the correct answer b (Figure 4). However, there was no association between answering question 6 incorrectly and making comparisons to absent controls on question 5, as determined by chi-square test. Thus, incomplete conceptions about experimental controls appeared to be widespread, but may involve different aspects of reasoning for designing and interpreting correctly controlled experiments.

3 | DISCUSSION

Learning the process of science is an essential part of any college education. In the biological sciences, forming testable hypotheses, designing experiments, and interpreting data should be integrated into the curriculum so that students are continually refining these scientific reasoning skills. The ability to think scientifically is not only necessary for obtaining a graduate degree or research-related career, it makes learning biology more fun and engaging by immersing students in the process of scientific discovery. An approach to teaching science that fosters the development of scientific reasoning along with learning content knowledge could improve student satisfaction, retention, and pursuit of STEM careers.[1,15,40] Thus, it is critical to effectively assess how students develop these skills in college.

Our assessment showed strengths and weaknesses in scientific reasoning for biology majors at our institution. Both freshman and senior-level students performed well at identifying dependent and independent variables for a testable hypothesis on question 1 (Table 2). Question 2 required higher-order Bloom's taxonomy skills and proved more challenging. To achieve the maximum score requires the evaluation Bloom's level to sort relevant from nonrelevant information and the synthesis Bloom's level to formulate a complete hypothesis.[29] Freshman students had more difficulty with this question but the senior-level students did not improve from pretest to posttest (Table 2). Question 2 was scored over a range of points based on the number of rubric items contained in the answer, and it was anticipated that it would challenge students' reasoning skills to get all the rubric items. Thus, the lower average scores on this question were not surprising, and the interpretation of the average score from one round of the assessment was only meaningful relative to another. Question 2 was more complex and bore greater scrutiny in the validation. In opinion-survey interviews, students rated the easier rubric items as at least somewhat likely to be included by a student who excelled at forming hypotheses, but rated the more difficult rubric items between somewhat likely and not very likely to be included. It is difficult to say how much the interviewees' familiarity and skill level with this type of reasoning weighed on their opinion. The in-depth validation interviews along with the other evidence of the question's functionality suggests that, as a whole, it had some value for determining more advanced student skills for hypothesis formation. One advantage to the free-response format is that the relative weighting of the rubric elements can easily be adjusted for future use.

The data interpretation questions (3a, 3b, and 4) varied in difficulty, and the average student scores across these questions reflected this. Most students did well interpreting the result of a time-course experiment from a Western blot, but determining the result of an experiment designed to show causation was more challenging (questions 3a and 3b). Likewise, using the concept of competition to interpret a graph on question 4 proved difficult, particularly for first-year students. To answer these data interpretation questions, students must carry out multiple cognitive actions that may include reading the graph or data representation, deciphering the experimental design to understand the variables and controls, and applying background information to interpret the result. Difficulty with one or more of the cognitive actions may affect student performance. This distinguishes these three questions from the other assessment questions, which require fewer distinct actions to answer successfully.

It is worth considering how the cognitive actions for interpreting the data representation, deciphering the experimental design, and applying background information affected students' abilities to answer the data interpretation questions. All three questions require students to read a data representation; the Western blot band patterns for questions 3a and 3b and an *x*, *y*-plot for question 4. Being fluent at reading data representations is essential for working in the biological sciences, but it is often taken for granted that students are proficient at this.[22] In the absence of instruction for interpreting various types of representations, undergraduates may struggle with correctly determining what they depict.[41,42] Thus, this is an integral part of measuring students' abilities to interpret experiments. For the experimental design, the complexity increases sequentially in questions 3a, 3b, and 4, and this factor could explain much of the difference in performance between different rounds of the assessment. The UD-Lab course has a Western blot experiment, although with a different design than the experiments in questions 3a and 3b. Western blot experience could help with these questions, but there was no significant difference between the pretest and posttest averages for these questions in the UD-Lab. The data interpretation questions also required some level of preexisting content knowledge (Table 1), here considered as distinct from knowledge of the experimental design. While we tried to minimize the amount of content knowledge needed, it cannot be excluded as a factor affecting question performance. This is particularly important when considering the difference in question scores between the UD-Lab seniors and the Intro-Lab freshmen.

The integration of these three cognitive actions is needed for all the data interpretation questions but is probably more important for question 4, which had the highest level of complexity. Graphs and figures that convey information about the external environment are external representations, whereas internal representations, or constructs of the mind, are cognitive actions carried out in relation to these external representations. Zhang and Norman described the integration of external and internal representations to form distributed representations that allow the performance of abstract tasks.[43] Distributed representations are mental models that result from the integration of different types of information, in this case reading a graph and applying background knowledge to interpret it. Schonborn and Anderson have further explained this for understanding external representations in biochemistry.[44] They defined understanding the mode in which the information is represented and the relevant discipline-specific concepts as coming together with reasoning ability to form mental models for what biochemical representations depict. Proficiency at forming these mental models may be an important difference between novice and expert-level reasoning ability.[22] Some of the other assessment questions of higher complexity may also require the integration of multiple cognitive actions to form mental models. Question 2 requires the synthesis of a hypothesis from a set of disparate facts. While this question does not require content knowledge or reading a representation, it may elicit the use of multiple cognitive actions to sort relevant from nonrelevant information and make meaningful interconnections. Although we cannot precisely define these cognitive actions, their integration in the form of a mental model may be necessary for a successful approach to the question.

Question 5 of the primary assessment measured students' skill at designing an experiment to test a given hypothesis. The rubric designated points for positive attributes of the experimental design, such as minimizing unwanted variables and selecting appropriate controls, and for making appropriate predictions about the experimental conditions. On average, students performed at a moderate level on this question and there was a trend toward improvement from freshman students in the Intro-Lab to the senior students in the UD-Lab, and from pretest to posttest in the UD-Lab. This suggests students are gaining a better understanding of what it means to design an experiment as they progress through their time at the university. As they are exposed to research experimentation in different classes and practice designing experiments, their conceptions of what goes into designing an experiment may become more sophisticated. However, most students failed to define experimental conditions in a quantitative fashion. From the validation interviews, it is not clear, if this omission represents a common error in minimizing unwanted variables or if the question did not successfully prompt students to utilize their understanding. It has been reported that college introductory biology students do not have good understanding of the importance of sample size and repeating experiments,[19] so quantifying experimental conditions could be a difficulty of a similar vein. Further investigation of this is needed.

Improvement was seen primarily for the more challenging assessment questions that had lower average scores. These scores improved from the Intro-Lab to the UD-Lab, from pretest to posttest in the UD-Lab, or both. This suggests that these challenging questions measure skills that students are learning at the university, whereas questions where the freshman group performed well may represent skills learned in their secondary education. The score distribution for the assessment totals suggested that the questions provided a level of difficulty that captured a range of scientific reasoning skill levels. With the UD-Lab pretest median score at 50%, the assessment was suited to detect modest changes in reasoning ability. Thus, the assessment appears to be a useful instrument for measuring improvement in scientific reasoning.

Assessment can also flag common errors and incomplete conceptions that can be addressed to help students better learn scientific reasoning. Woolley et al. identified faulty reasoning patterns in undergraduate students for several defined scientific reasoning tasks such as designing experiments and building and interpreting graphs, although the prevalence of these reasoning errors in undergraduate populations was not reported.[45] We identified two potential widespread reasoning errors. The first occurred with high frequency on question 2, where students gave a hypothesis that was a restatement of the observation it was supposed to explain (Table 3). A surprising 38.3% of Intro-Lab freshmen answered incorrectly in this way. It is important to consider the cognitive source of this error, and whether this represents a common incomplete conception regarding what a hypothesis is. One possibility is that students misinterpreted what the question was asking them to do. Answering question 1 required simply restating information given in the question in an appropriate way, and considering the amount of information contained in question 2, it is possible students thought they needed to answer by extracting the correct phrase. However, the validation interviews indicated that students generally understood the question. One student who participated in an in-depth interview answered by restating the hypothesis but appeared to understand the question in the way it was intended. A limitation to the validation of question 2 is that in-depth interviews were conducted with only three students. Assuming the question is working correctly, it is possible this error represents a student's belief that a hypothesis can be a formal statement of something that is already known rather than a question about something that is not known. With this alternate conception, a student might propose a hypothesis as a restatement of an observation that had already been established, and not a testable statement that forms the basis for finding new information. A contributing factor to this error may have been the large amount of information that students had to sort in answering question 2. Cognitive overload could have increased the likelihood that a student would give the simplest answer possible, even if incorrect. The ability to sort relevant from nonrelevant information, though, is part of the reasoning experts carry out in conducting scientific investigation. This reasoning error was less frequent in the senior-level UD-Lab students than the freshman students, and less frequent on the UD-Lab posttest than the pretest.

The second widespread reasoning error occurred on question 5. Students frequently made comparisons between the samples they chose to run in the experiment and control samples that they did not include. The highest frequency of this error occurred in the freshman Intro-Lab (22.8%), and the frequency decreased with the senior-level UD-Lab students, and from pretest to posttest in the UD-Lab. Even on the UD-Lab posttest, this error occurred on 13.9% of the assessments. This suggests that students did not understand that control conditions need to be run in the experiment to validate that it is working as expected. Shi et al. reported a similar lack of understanding of controls, finding that undergraduates frequently indicated positive and negative control conditions were unnecessary for an experiment.[24]

The frequent comparisons to absent controls on the primary assessment struck the authors as similar to mistakes commonly observed with the UD-Lab students. When selecting sample conditions for a Western blot experiment to test how different reagents might affect growth factor signaling, students would sometimes leave out a “no additions” control, which is necessary to establish that the growth factor was stimulating the cells as expected. These errors may stem from an incomplete conception that previously determined experimental information creates a fixed point of reference to which future experiments can be compared, without repeating those conditions. Wooley et al. found that reliance on prior knowledge was a faulty strategy that pervaded many types of scientific reasoning.[45] In this case, the prior knowledge that a growth factor stimulates cells or that bean beetles have shown an egg-laying preference for one bean over another would prevent students from repeating these conditions as internal controls for an experiment. This creates the faulty assumption that the experiment must work correctly.

The secondary assessment further probed students' understanding of how controls are used in designing and interpreting experiments, and included a new question that posed a problem similar to what students struggled with in the UD-Lab (question 6). When asked to identify the experiment that provided the best evidence for how a potential inhibitor affected growth factor stimulation, 47% of students ($n = 291$) selected an experiment that was missing the critical “no additions” control (Figure 4), similar to the mistake observed in the UD-Lab. Question 6 asked students to demonstrate an understanding of controls in interpreting experimental data, while question 5 required them to use this understanding in designing an experiment. We had assumed that these were essentially the same cognitive action, and that making comparisons to absent controls on question 5 was the same reasoning error as selecting the uncontrolled experiment in question 6. However, there was no association between these errors on the secondary assessment. Students who identified the correctly controlled experiment on question 6 were as likely to make comparisons to absent controls on question 5 as students who answered question 6 incorrectly. Thus, the reasoning and specific cognitive actions for the use of controls may be complex and multifaceted. In any case, these errors may pose a roadblock for students learning scientific reasoning. Including instruction on experimental controls has been shown to improve student understanding of their use.[24]

The ability to think critically is the most important attribute of a college education. As educators, it is incumbent on us to ensure our undergraduates are achieving this goal. Instruction of scientific reasoning should be implemented throughout a biological sciences undergraduate curriculum. Students best learn when these skills are tailored to specific subject matter, so their instruction should be threaded into the content of all biology courses. Assessing scientific reasoning is an inseparable component of teaching it. The greater the number of assessments available to instructors, and the more varied they are in discipline-specific content, the easier it will be to achieve these learning outcomes. By adopting a curricular strategy that cycles instruction followed by assessment to inform further instruction, we will best be able to foster these skills in our students.

ACKNOWLEDGMENTS

The authors thank Lisa McDonnel for helpful discussions on forming mental models in scientific reasoning, and Kesten Bozinovic for contributions to scoring the assessments. The research in this paper was not supported by grant funding.

ORCID

Aaron B. Coleman https://orcid.org/0000-0002-9198-1451
Goran Bozinovic https://orcid.org/0000-0001-8934-7884

REFERENCES

1. Brewer CA, Smith D, editors. Vision and change in undergraduate biology education: a call to action. Washington, DC: AAAS; 2011.
2. Coil D, Wenderoth MP, Cunningham M, Dirks C. Teaching the process of science: faculty perceptions and effective methodology. CBE Life Sci Educ. 2010;9(4):524–35. https://doi.org/10.1187/cbe.10-01-0005
3. Zimmerman C. The development of scientific reasoning skills. Dev Rev. 2000;20(1):99–149. https://doi.org/10.1006/drev.1999.0497
4. Blumer LS, Beck CW. Laboratory courses with guided-inquiry modules improve scientific reasoning and experimental design skills for the least-prepared undergraduate students. CBE Life Sci Educ. 2019;18(1):ar2. https://doi.org/10.1187/cbe.18-08-0152
5. Narayan R, Rodriguez C, Araujo J, Shaqlaih A, Moss G. Constructivism: constructivist learning theory. In: Irby BJ, Brown G, Lara-Alecio R, Jackson S, editors. The handbook of educational theories. Charlotte, N.C: IAP Information Age Publishing; 2013. p. 169–83.
6. Hammer D. Student resources for learning introductory physics. Am J Phys. 2000;68(7):S52–9. https://doi.org/10.1119/1.19520
7. Cotner S, Ballen CJ. Can mixed assessment methods make biology classes more equitable? PLoS One. 2017;12:e0189610. https://doi.org/10.1371/journal.pone.0189610
8. Maskiewicz AC, Lineback JE. Misconceptions are "so yesterday!". CBE Life Sci Educ. 2013;12(3):352–6. https://doi.org/10.1187/cbe.13-01-0014
9. Killpack TL, Fulmer SM. Development of a tool to assess interrelated experimental Design in Introductory Biology. J Microbiol Biol Educ. 2018;19(3):1627–1637. https://doi.org/10.1128/jmbe.v19i3.1627
10. Dasgupta AP, Anderson TR, Pelaez N. Development and validation of a rubric for diagnosing students' experimental design knowledge and difficulties. CBE Life Sci Educ. 2014;13(2):265–84. https://doi.org/10.1187/cbe.13-09-0192
11. Smith KA, Welliver PW. The development of a science process assessment for fourth-grade students. J Res Sci Teach. 1990;27(8):727–38. https://doi.org/10.1002/tea.3660270803
12. Shahali EHM, Halin L. Development and validation of a test of integrated science process skills. Procedia Soc Behav Sci. 2010;9:142–6. https://doi.org/10.1016/j.sbspro.2010.12.127
13. Dillashaw FG, Okey JR. Test of the integrated science process skills for secondary science students. Sci Educ. 1980;64(5):601–8. https://doi.org/10.1002/sce.3730640506
14. Burns JC, Okey JR, Wise KC. Development of an integrated process skill test: TIPS II. J Res Sci Teach. 1985;22(2):169–77. https://doi.org/10.1002/tea.3660220208
15. Dirks C, Cunningham M. Enhancing diversity in science: is teaching science process skills the answer? CBE Life Sci Educ. 2006;5(3):218–26. https://doi.org/10.1187/cbe.05-10-0121
16. Kramer M, Olson D, Walker JD. Design and assessment of online, interactive tutorials that teach science process skills. CBE Life Sci Educ. 2018;17(2):1–11. https://doi.org/10.1187/cbe.17-06-0109
17. Tobin KG, Capie W. Development and validation of a group test of integrated science processes. J Res Sci Teach. 1982;19(2):133–41. https://doi.org/10.1002/tea.3660190205
18. Sirum K, Humburg J. The experimental design ability test (EDAT). Bioscene: J College Biol Teach. 2011;37(1):8–16.
19. Brownell SE, Wenderoth MP, Theobald R, Okoroafor N, Koval M, Freeman S, et al. How students think about experimental design: novel conceptions revealed by in-class activities. Bioscience. 2014;64(2):125–37. https://doi.org/10.1093/biosci/bit016

20. Dean T, Nomme K, Jeffery E, Pollock C, Birol G. Development of the biological experimental design concept inventory (BEDCI). CBE Life Sci Educ. 2014;13(3):540–51. https://doi.org/10.1187/cbe.13-11-0218
21. Rybarczyk BJ, Walton KLW, Grillo WH. The development and implementation of an instrument to assess students' data analysis skills in molecular biology. J Microbiol Biol Educ. 2014; 15(2):259–67. https://doi.org/10.1128/jmbe.v15i2.703
22. Schonborn KJ, Anderson TR. The importance of visual literacy in the education of biochemists. Biochem Mol Biol Educ. 2006; 34(2):94–102. https://doi.org/10.1002/bmb.2006.49403402094
23. Dasgupta AP, Anderson TR, Pelaez NJ. Development of the neuron assessment for measuring biology students' use of experimental design concepts and representations. CBE Life Sci Educ. 2016;15(2):1–21. https://doi.org/10.1187/cbe.15-03-0077
24. Shi J, Power JM, Klymkowsky MW. Revealing student thinking about experimental design and the roles of controls in experiments. Int J Scholarship Teach Learn. 2011;5(2):8. https://doi.org/10.20429/ijsotl.2011.050208
25. Coleman AB, Lam DP, Soowal LN. Correlation, necessity, and sufficiency: common errors in the scientific reasoning of undergraduate students for interpreting experiments. Biochem Mol Biol Educ. 2015;43(5):305–15. https://doi.org/10.1002/bmb.20879
26. Stanhope L, Ziegler L, Haque T, Le L, Vinces M, Davis GK, et al. Development of a biological science quantitative reasoning exam (BioSQuaRE). CBE Life Sci Educ. 2017;16(4):ar66. https://doi.org/10.1187/cbe.16-10-0301
27. Gormally C, Brickman P, Lutz M. Developing a test of scientific literacy skills (TOSLS): measuring undergraduates' evaluation of scientific information and arguments. CBE Life Sci Educ. 2012;11(4):364–77. https://doi.org/10.1187/cbe.12-03-0026
28. Zhou S, Han J, Koenig K, Raplinger A, Pi Y, Li D, et al. Assessment of scientific reasoning: the effects of task context, data, and design on student reasoning in control of variables. Think Skills Creat. 2016;19:175–87. https://doi.org/10.1016/j.tsc.2015.11.004
29. Bloom BS. Taxonomy of educational objectives: the classification of educational goals—handbook 1, cognitive domain. New York, NY: David McKay; 1956.
30. Stanger-Hall KF. Multiple-choice exams: an obstacle for higher-level thinking in introductory science classes. CBE Life Sci Educ. 2012;11(3):294–306. https://doi.org/10.1187/cbe.11-11-0100
31. Beggrow EP, Ha M, Nehm RH, Pearl D, Boone WJ. Assessing scientific practices using machine-learning methods: how closely do they match clinical interview performance? J Sci Educ Technol. 2014;23:160–82. https://doi.org/10.1007/s10956-013-9461-9
32. Scalise K, Gifford B. Computer-based assessment in E-learning: a framework for constructing "intermediate constraint" questions and tasks for technology platforms. J Technol Learn Assess. 2006;4(6). https://ejournals.bc.edu/index.php/jtla/article/view/1653
33. Meira E, Wendel D, Pope DS, Hsiao L, Chen D, Kim KJ. Are intermediate constraint question formats useful for evaluating student thinking and promoting learning in formative assessments? Comput Educ. 2019;141:103606. https://doi.org/10.1016/j.compedu.2019.103606
34. Treagust DF. Development and use of diagnostic tests to evaluate students' misconceptions in science. Int J Sci Educ. 1988; 10(2):159–69. https://doi.org/10.1080/0950069880100204
35. Corbin JM, Strauss A. Grounded theory research: procedures, cannons, and evaluative criteria. Qual Sociol. 1990;13(1):3–21. https://doi.org/10.1007/BF00988593
36. Russell CB, Weaver G. Student perceptions of the purpose and function of the laboratory in science: a grounded theory study. Int J Scholarship Teach Learn. 2008;2(2):9. https://doi.org/10.20429/ijsotl.2008.020209
37. Jones P, Lucock M, Veysey M, Beckett E. The vitamin D-folate hypothesis as an evolutionary model for skin pigmentation: an update and integration of current ideas. Nutrients. 2018;10(5): 554. https://doi.org/10.3390/nu10050554
38. Pierschbacher MD, Ruoslahti E. Cell attachment activity of fibronectin can be duplicated by small synthetic fragments of the molecule. Nature. 1984;309:30–3. https://doi.org/10.1038/309030a0
39. Beck CW, Blumer LS, Habib J. Effects of evolutionary history on adaptation in bean beetles, a model system for inquiry-based laboratories. Evo Educ Outreach. 2013;6:article 5. https://doi.org/10.1186/1936-6434-6-5
40. Seymour E, Hewitt NM. Talking about leaving: why undergraduates leave the sciences. Boulder, CO: Westview Press; 1997.
41. Bowen GM, Roth W, McGinn MK. Interpretations of graphs by university biology students and practicing scientists: toward a social practice view of scientific representation practices. J Res Sci Teach. 1999;36(9):1020–43. https://doi.org/10.1002/(SICI)1098-2736(199911)36:9<1020::AID-TEA4>3.0.CO;2-%23
42. Bowen GM, Roth W. Why students may not learn to interpret scientific inscriptions. Res Sci Educ. 2002;32:303–27. https://doi.org/10.1023/A:1020833231966
43. Zhang J, Norman DA. Representations in distributed cognitive tasks. Cogn Sci. 1994;18(1):87–122. https://doi.org/10.1207/s15516709cog1801_3
44. Schonborn KJ, Anderson TR. A model of factors determining students' ability to interpret external representations in biochemistry. Int J Sci Educ. 2009;31(2):193–232. https://doi.org/10.1080/09500690701670535
45. Woolley JS, Deal AM, Green J, Hathenbruck F, Kurtz SA, Park TKH, et al. Undergraduate students demonstrate common false scientific reasoning strategies. Think Skills Creat. 2018;27: 101–13. https://doi.org/10.1016/j.tsc.2017.12.004

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Coleman AB, Lorenzo K, McLamb F, Sanku A, Khan S, Bozinovic G. Imagining, designing, and interpreting experiments: Using quantitative assessment to improve instruction in scientific reasoning. Biochem Mol Biol Educ. 2023. https://doi.org/10.1002/bmb.21727