Motivating and Shaping Scientific Argumentation in Lab Reports

Julia Gouvea,[†‡]* Lara Appleby,[†§] Liren Fu,[‖] and Aditi Wagh[¶]

[†]Department of Education, [‡]Department of Biology, and [§]Department of Physics and Astronomy, Tufts University, Medford, MA 02155; [‖]Ministry of Education, Singapore 138675; [¶]Comparative Media Studies and Writing, Massachusetts Institute of Technology, Cambridge, MA 02142

ABSTRACT

Writing a lab report can be an opportunity for students to engage in scientific thinking. Yet students' lab reports often do not exhibit evidence of such engagement. Students' writing can appear focused on "filling in" required components and reporting on predetermined conclusions. We conducted a design experiment in an introductory biology laboratory course and examined the impact on students' engagement in argumentation in lab reports. Over two design iterations, students' arguments more often considered and integrated multiple claims, included a broader range of evidence and ideas, and gave appropriate attention to uncertainty in conclusions. We argue that two interrelated changes to the design of the lab course made these shifts possible. First, we restructured the role of instructors to position them as an audience interested in students' thinking. Second, we introduced more uncertainty into the lab activities to provoke consideration of multiple interpretations. We propose that these changes created a different *rhetorical context* that helped motivate and shape students' engagement in argumentation. More broadly, we suggest that an important alternative to explicitly scaffolding knowledge and skills is to design learning environments that can inspire students to engage in a range of scientific practices more authentically.

INTRODUCTION

Writing is a process of scientific thinking, not just a means of recording scientific findings. Its capacity to engage creative and critical thought is one of the main reasons scientific writing has been promoted as a central activity across the science curriculum (Quitadamo and Kurtz, 2007; Libarkin and Ording, 2012; Walker and Sampson, 2013). In the undergraduate curriculum, one of the most common forms of writing is the lab report. Research on lab report writing has revealed what many instructors recognize from experience—lab reports often fail to elicit students' scientific thinking (e.g., Moskovitz and Kellogg, 2011). Instead, lab report assignments can promote writing that contains formulaic arguments that report on predetermined conclusions (Keys, 1999; Xu and Talanquer, 2013) and includes evidence without reasoning or justification (Kelly and Bazerman, 2003; Schen, 2017).

We report on a research study in an introductory-level biology laboratory course born out of observations of writing assignments not living up to their potential to elicit scientific thinking. We observed, for example, that many students' lab reports were organized around demonstrating a fact or principle from the lab manual, even when those principles could not be plausibly demonstrated by the data. Students appeared to be organizing their reports around the conclusions they expected. In addition, reports included analyses of sources of error, but rarely did such analyses function to alter the certainty of conclusions. These observations suggested that, rather than authentically engaging in scientific argumentation, students may have been more focused on meeting perceived expectations of lab report writing. In the science education literature, this phenomenon has been

Vicente Talanquer, *Monitoring Editor*

Submitted Nov 4, 2021; Revised Aug 18, 2022; Accepted Aug 31, 2022

CBE Life Sci Educ December 1, 2022 21:ar71

DOI:10.1187/cbe.21-11-0316

*Address correspondence to: Julia Gouvea (julia.gouvea@tufts.edu).

referred to as "pseudoargumentation" to emphasize that, while students' writing or speech may contain features of formal scientific arguments, those features are not serving an authentic rhetorical function (Kelly *et al.*, 2010; Berland and Hammer, 2012; Manz, 2015a).

Pseudoargumentation limits opportunities for students to engage in critical or creative thought during the process of writing. Interpreting data through the lens of expected claims from a lab manual, students have no good reason to consider and compare alternative interpretations of their data (Jiménez-Aleixandre *et al.*, 2000; Manz *et al.*, 2020). When engaged in pseudoargumentation, students may present claims with a high degree of certainty, explaining away deviations from expectations through acknowledgments of "error." These errors are presented as evidence of mistakes in technique rather than a reason to consider adjusting claims (Holmes and Bonn, 2014; Lee *et al.*, 2014; Stein *et al.*, 2018). Such moves undermine critical considerations of the nature and limits of scientific knowledge. Finally, while pseudoargumentation may allow students to learn to *reproduce* argumentative forms such as claim statements or evaluations of sources of error, students do not gain practice deciding how and when to *use* rhetorical elements to build better arguments (Ford, 2012; Manz, 2015a).

In this paper, we describe a design experiment (Cobb *et al.*, 2003) aimed at shifting students' engagement in argumentation in lab reports. The main purpose of design experiments, and design-based research more broadly, is to generate and refine theories about *how* designed features of learning environments support learning outcomes (Cobb *et al.*, 2003; Sandoval, 2014). In this work, we examined the links between the designed features of an introductory biology laboratory course and how students engaged in argumentation in their lab reports.

Our approach to design was informed by research in K–12 science education, showing that explicit instruction and highly structured scaffolding can reinforce rote engagement in scientific practices (Jiménez-Aleixandre *et al.*, 2000; Ford, 2005; Berland and Hammer, 2012; Russ and Berland, 2019). As Manz (2015a) argues in her review of argumentation in science education, approaches that aim to teach students how to properly structure an argument can inadvertently lead to pseudoargumentation, as students focus on following directions and anticipating the teachers' desired conclusions instead of trying to express their thinking. This can occur because the learning environment influences students' understandings, or *framings*, of what kind of behavior is valued or appropriate (Jiménez-Aleixandre *et al.*, 2000; Hammer *et al.*, 2005; Scherr and Hammer, 2009; Berland and Hammer, 2012; Petritis *et al.*, 2021). A focus on formal rules makes it more likely that students frame their role as complying with these rules. Thus, in our design, we did not provide direct, explicit support for the kind of behavior we wanted but rather attempted to change the learning environment in ways that would make it more likely that authentic forms of argumentation would emerge (see also Zagallo *et al.*, 2016). In short, we worked to design a learning environment that would help students frame writing a lab report as an activity of expressing their thinking. We refer to the target of our design as the *rhetorical context* of the lab report assignment—students' understanding of both why they are writing and to whom. We claim that, by designing to shift the rhetorical context, we were able to support students' engagement in more authentic forms of scientific argumentation that include evidence of critical and creative thought.

In the next section, we describe the importance of rhetorical context in framing argumentation practices in scientific communities. We then review research that has demonstrated how learning environments contribute to science learners' framings of the task of argumentation. After this background, we describe our own design experiment and present evidence of shifting engagement in argumentation over the two design iterations of our study.

The Role of Context in Scientific Argumentation

Scientific articles did not always include arguments. Bazerman's (1988) analysis of early physics journals found that, before the 1930s, the majority of articles simply reported on experimental findings. Arguments that included reasoning to support data interpretations began appearing in scientific writing as the community began to change its understanding of the scientific enterprise. Physicists began to see the natural world not as something to be observed and written down, but as puzzling and a "matter of contention" (Bazerman, 1988, p. 78). As the community began to grapple with the understanding that observations could be interpreted in more than one way, authors began including sections devoted to "discussion." In these sections, authors attempted to persuade readers of particular interpretations of their data over alternatives.

While Bazerman's (1988) account is focused on the physics community, we take the following general points from his analysis: 1) argumentation arises in scientific communities in response to *problems*: unexplained phenomena, unsettled debates, or unnoticed inconsistencies; and 2) arguments are crafted in anticipation of a response from a critical *audience* of peers who will evaluate the quality of the argument according to the evolving criteria valued by the community. These features of context—the problems and audience—have been recognized as motivating and shaping rhetorical products generally across many different domains (Bitzer, 1968; Bazerman, 1988, 2018; Petraglia, 1995). Scholars of rhetoric have argued that it is not just the *skills* of an author (or orator) that contribute to a rhetorical product, but the "*situation* which invites the orator's application of his method and the creation of discourse" (Bitzer, 1968, p. 2, emphasis added).

This view aligns with recent understandings of argumentation as a scientific practice that emerges in response to a felt "need" (Berland and Reiser, 2011; Berland and Hammer, 2012; Manz, 2015a,b; Chen *et al.*, 2019). Without a context to inspire a need to solve a problem or convince an audience, skills and knowledge relevant to argumentation can lay dormant. In this view, learning to argue like a scientist requires opportunities to experience contexts that motivate and shape argumentation. Yet many science learning environments lack the features of context that are important for stimulating argumentation. This has led researchers to propose that perhaps the contexts of school argumentation, more so than a lack of student ability, can account for the apparent lack of sophistication among novices (Keys, 1999; O'Neill, 2001; Ford, 2012; Manz, 2015a).

The Role of Context in Framing Student Argumentation

One strand of research on scientific argumentation has focused on students' lack of sophistication: students' failure to provide

sufficient evidence for claims, inadequate explanations, and lack of attention to counterarguments or flaws in their arguments (Lawson, 2002; Kuhn and Udell, 2003; Osborne *et al.*, 2004, 2016). These patterns have been explained in terms of novices' lack of knowledge and skills. Researchers have responded by designing and studying interventions to support skill development (e.g., Osborne *et al.*, 2004; Sandoval and Millwood, 2005; McNeill *et al.*, 2006). While these efforts have produced some evidence of improvement in understanding specific skills, providing more instruction and support has not often resulted in more sophisticated argumentation (Manz, 2015a). Students' arguments include target features, but often still lack evidence of purposeful application of strategies or critical engagement with the ideas.

For example, Sandoval and Millwood (2005) studied high school students' use of a software tool designed to scaffold argumentation by prompting students with text boxes that could be used to link evidence to claims. While the tool was successful in getting students to use data as evidence, the majority of students simply included links to evidence without explaining how it supported their claims. The authors suggest that this may have been because students did not perceive themselves to be arguing to an authentic audience. As the authors point out, the whole class, including the teacher, had access to the same data set. Therefore, the authors conjecture, "It is possible, even likely, that students perceived the persuasive goal of their explanations to be to show the teacher that they had figured out the right answer" (Sandoval and Millwood, 2005, p. 49). In this context, appending shared evidence could be understood as sufficient to satisfy the demands of the task. What this and other research has shown is that explicit instruction and scaffolding can support rote skill application, such as appending evidence, but do not necessarily lead to flexible or purposeful applications, such as deciding which evidence to use and how to use it (Kuhn and Pease, 2008; Ford, 2012; Manz, 2015a).

Classroom context can influence how students engage in scientific practices through an effect on students' framings of tasks and activities (Jiménez-Aleixandre *et al.*, 2000; Berland and Hammer, 2012). Framing theory has roots in anthropology and socio-linguistics where it has been used to account for how people make sense of "what is going on" in a situation (Tannen, 1979; Tannen and Wallat, 1987; Goffman, 1986). The process of framing is described as an interaction between an individual's prior knowledge and experiences and cues from the current context. A classic thought experiment describes how framing applies to the situation of entering an unfamiliar restaurant (Schank and Abelson, 1977). One might notice the physical arrangement of tables; the presence of a host; the brightness of the lighting; the volume of conversations; and other physical, social, and cultural cues. These cues are interpreted through one's prior knowledge and experiences, helping one to decide what kind of place this is and to adjust one's behavior accordingly—seating oneself or waiting to be seated, for example. This same process occurs when students enter learning spaces: The physical arrangement of the space and the social, cultural, and epistemic cues interact with students' prior experiences to form interpretations of what is expected and valued. While students are active participants in the framing process, the possible framings can be heavily constrained by how the learning environment has been designed.

Applied to argumentation, framing has been used to explain more authentic versus more rote engagement in argumentative discourse and writing (Jiménez-Aleixandre *et al.*, 2000; Berland and Reiser, 2011; Berland and Hammer, 2012; Petritis *et al.*, 2021). Jiménez-Aleixandre *et al.* (2000) proposed that two contrasting framings accounted for shifts in high school students' conversations about genetics. In one framing—"doing the lesson"—students seemed to understand their role as conforming to a set of expectations of how to be a "good student." This meant constructing arguments that matched expected content (Jiménez-Aleixandre *et al.*, 2000, p. 770), a focus that left little room for students to do their own thinking, engage with evidence, or attend to one another's ideas. Later in the lesson, the authors identify a switch in the dialogue indicating that the class had constructed a different framing of the conversation, more in line with "doing science," in that students seemed to understand their role as legitimately trying to "figure things out" for themselves. Students proposed and defended their ideas, used analogies to explain their thinking, and evaluated one another's ideas by making appeals to consistency.

The authors explain the shifts in framing in terms of features of students' incoming expectations and aspects of the classroom context. "Doing the lesson" seemed to emerge in part due to the prevalence of a common set of expectations about how students are supposed to behave in school. Students often expect that they are meant to comply with instructions and demonstrate that they have learned what the teacher intended to teach them. The authors suggest that these expectations may have been cued up and stabilized by a worksheet that kept students focused on filling in answers. "Doing science" emerged when legitimate disagreements about alternative explanations for the phenomenon arose within and among student groups. Students shifted their attention to the disagreement itself, and argumentation began to function to help them to better understand different ideas. Jiménez-Aleixandre *et al.* (2000) further explain that students' willingness to engage with the disagreements was made possible by an existing local classroom culture in which students felt comfortable voicing their own ideas and questions. These features of context made it possible for students to shift into a pattern of arguing together.

Drawing on this and other studies, Manz (2015a) argued that authentic argumentation is more likely to emerge when students encounter something problematic that cannot be easily resolved. When there is a problem, argumentation can function as a tool to help students identify and evaluate different ideas. In addition, such conversations require a shared understanding of the social context as one in which students can freely engage with one another rather than appeal only to the teacher. That is, the rhetorical context that can elicit argumentation in classrooms resembles two important features of scientific contexts—the existence of legitimate problems and communication with an authentic audience.

We used these ideas to inform our redesign of an introductory biology lab course by altering the structure of instruction and assessment and by increasing the uncertainty students experienced as they analyzed data. We next describe these changes in more detail and then explain the methods that we used to characterize students' engagement in argumentation in

TABLE 1. Summary of laboratory course design from 2014–2016.

Design features	2014	2015	2016
Background and lab purpose	Background on *E. coli* strains Purpose of lab is to "show that mutations can sometimes be beneficial"	Background on *E. coli* strains Purpose of lab to investigate the question: Is it better to mutate a lot or a little (and under what conditions)?	
Lab activities: experiment	Basic experimental procedure including plating on three types of media (LB agar, LB + lac, MacConkey) shows advantage to mutating in novel environments	Plating on LB agar and LB + rif only. Colonies on LB agar are countable allowing inferences about fitness in nonselective media Competition experiment comparing strains plated on LB media supports multiple predictions and outcomes	
Lab activities: computer simulation	None	Thirty-minute agent-based simulation activity to explore conditions under which virtual strains with different mutation rates have advantage	One-hour simulation activity allows students to run comparisons and plots to show mutation frequencies GTAs emphasize "puzzling" output for whole class
Role of instructors	Pre-lab quiz GTA lecture reviewing lab manual Optional discussion questions	Small-group and whole-class discussions about relative benefits of higher/lower rate of mutation Small-group and whole-class discussions of patterns and possible interpretations	
Assessment guidelines	Guidelines emphasize following directions Discussion section should include a statement of whether data conformed to the expected hypothesis	Guidelines emphasize students making sense of data Discussion section should discuss "ideas and evidence" Optional inclusion of output from simulation	
Peer review	Draft and exchange methods in class	Peer review (introduction, methods, results only)	Peer review of discussions emphasizing looking for consistent reasoning

lab reports to track what, if any, shifts accompanied the course redesign.

METHODS

Study Context

The context for this study was a semester-long introductory laboratory course taught at Tufts University, a small, private, research-intensive liberal arts university. The course typically enrolled about 350–400 students, the majority of whom were freshmen and sophomores who had not yet declared a major. The course was required for biology majors, but students could choose to use Advanced Placement credit to place out of this requirement. Students simultaneously enrolled in both the lecture and laboratory components of the course. The laboratory component accounted for 25% of students' final grade, and the lab grade was largely determined by lab report scores. Over a semester, students attended a series of nine 3-hour lab classes led by graduate student teaching assistants (GTAs). GTAs were trained by a faculty laboratory coordinator during a 3-hour preparatory session each week. In lab, students worked in groups of three or four, but all lab reports were written and graded individually.

We compared lab reports written for the first unit of the semester over three successive years. The focus of this unit was an investigation of two different strains of the bacterium *Escherichia coli.* One strain was a typical laboratory strain, and the second strain had a dysfunctional DNA repair system that caused it to mutate at a higher rate than the wild-type strain. The central phenomenon concerned the relative fitness of these two phenotypes (higher and lower rates of mutation). Mutations can damage functional DNA, reducing fitness. However, mutations are also a source of variation, enabling populations to adapt to novel conditions.

Design Iterations

We analyzed lab reports over a period of 3 years: the original lab course before any design intervention (2014) and two iterations of redesign (2015 and 2016). Because this time period represented the beginning of a redesign effort, the changes to the curriculum and instruction were guided both by design aims and practical constraints. In this section, we describe the design of the lab over the 3 years that we collected data. While we strive to be comprehensive in our documentation of design changes, our descriptions were generated through a retrospective analysis of design choices likely to influence students' framing of lab report writing. To generate design descriptions, we reviewed curricular and instructional materials, including the syllabus, lab manual and in-class handouts, notes and slides prepared for GTAs, and grading guidelines and rubrics for lab report assignments. In Table 1, we summarize key features of the design iterations in terms of 1) how materials communicated the background and purpose of the laboratory unit, 2) the activities (experiments and computer simulations), 3) the role of instructors, 4) and the purpose and structure of assessments (by instructors and peers).

Across the three versions of the unit, students worked with the same basic experimental system, comparing survival and growth of two strains of *E. coli* on several types of growth

media. One strain (E938) mutates at a roughly 100-fold higher rate than the other (E939). Neither strain was resistant to the antibiotic rifampicin (i.e., the strains were both initially rif–) and neither could initially digest lactose (i.e., the strains were both initially lac–). Before class, the two strains of *E. coli* were incubated overnight from a frozen stock. Students could plate the strains on three different media: a nutrient-rich medium (LB agar), a nutrient-rich medium with the antibiotic rifampicin added (LB + rif), and a nutrient-rich medium with lactose added and an indicator dye that turns red when lactose is metabolized (LB + lac), also called MacConkey media. After a few days, students assessed cell growth by counting colonies that were visible on the plates. Colony number was impacted by both the concentration of cells plated and the selectivity of the media. Typical results are shown in Figure 1. A version of this experiment was conducted in all 3 years, though some details changed, and some activities were added, as we describe in the below sections.

Design of the Original Lab. The original curriculum had many features common among traditional introductory laboratory courses. The lab manual presented background information on the study system and introduced the purpose of the experiments as follows: "Most mutations are deleterious, but very rarely, a mutation can cause an increase in fitness. We can see how this takes place by conducting an experiment." This phrasing suggested that the experiments were intended to demonstrate how mutations can increase fitness. The manual also provided a description of the experimental procedures and included a data sheet into which students could enter their data. Students were expected to read the lab manual before arriving in the lab, an expectation that was enforced with a quiz administered at the start of the lab unit.

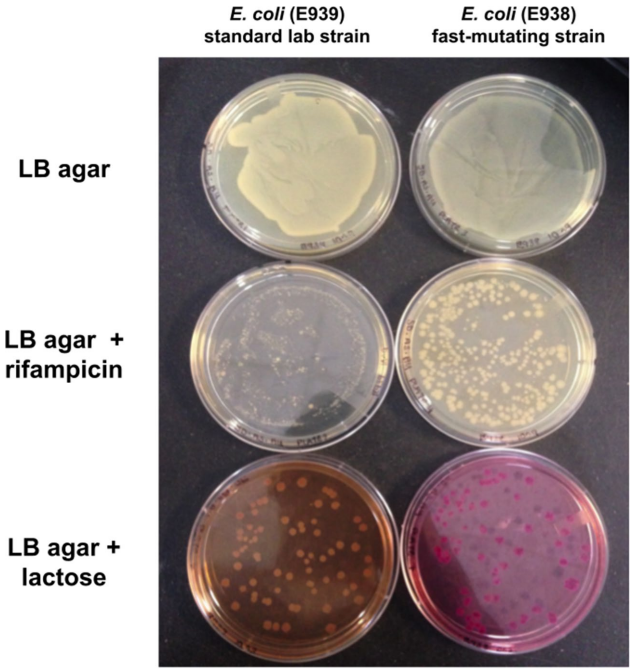

FIGURE 1. Typical experimental results from 2014. On nutrient-rich LB agar, both strains grow many colonies. In this example, the colonies have fused into a lawn due to a high concentration of cells plated initially. On LB + rif media, more colonies of the fast-mutating strain will survive, indicating a larger proportion of cells that have an antibiotic resistance mutation. On the LB + lac plates, overall numbers of colonies are similar on average, but the fast-mutating strain has more colonies that are red, indicating that these colonies have a mutation allowing them to digest lactose.

Instruction by GTAs was designed primarily to emphasize core concepts as well as to guide students through the details of implementing experimental procedures. At the beginning of class, the GTA presented slides that reviewed the main points in the lab manual, including key concepts such as "mutations are the ultimate source of genetic variation," once again emphasizing the intended conceptual takeaway of the lab. The GTA also presented an overview of the experimental procedures, highlighting important steps, common pitfalls, and safety considerations. Students then implemented the basic version of the experiment as described earlier, working in groups of three or four while the instructor circulated the room to answer questions and provide guidance. If students completed the lab protocol correctly, they typically found results showing more antibiotic-resistant and lactose-digesting colonies for the faster-mutating strain (Figure 1).

Instructions for writing the lab report were described in the lab manual (see Appendix A in the Supplemental Material), beginning with general advice that emphasized following directions: "To do well, you must follow directions and work ahead so that you have time to proof-read your report before handing it in to your instructor." The guidelines further elaborated on what should be in each of the lab report sections (i.e., methods, results, discussion). Reports were graded using a 20-item rubric that specified the breakdown of point values for each section of the lab report. An analysis of slides created for GTAs indicated that the GTAs highlighted key aspects of the rubric in presenting the lab report assignment. Specifically, the slides outlined that the discussion section should include "a one-sentence summary of what you found; a statement of how your results conform to your hypothesis; if your results are unexpected, offer some explanations; and suggestions for future experiments that should be conducted."

In addition, at the end of the unit, students were prompted to draft a methods section and were given class-time to exchange the draft with a peer for review.

First Design Iteration. In the first redesign of this unit, we changed the lab manual so that it provided background information on the *E. coli* strains (e.g., genotypes of two strains and a description of the DNA repair system and the mutation that caused dysfunction in DNA repair), but we removed references to expected interpretations. Instead, the pre-lab readings introduced the phenomenon that mutation rate is a variable trait in natural populations and proposed as a guiding question for the unit: "Is it better to mutate a lot or a little (and under what conditions)?"

Another major change in this iteration was to restructure the role of the GTA instructors (Table 1). In place of an introductory lecture, GTAs led a series of discussions intended to provoke ideas about why mutation rates might vary across and within species. For example, one question asked students to predict whether *E. coli*, *Arabidopsis thaliana*, or *Homo sapiens* would be

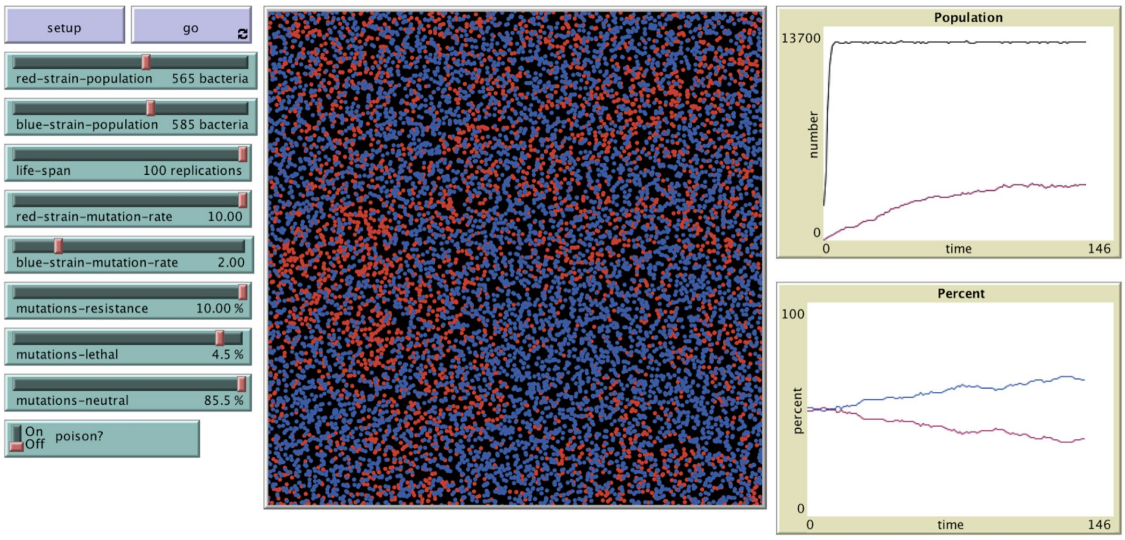

FIGURE 2. Simulation interface and example output (2015). Sliders allow students to change initial population sizes of two strains, the number of times a cell replicates before it dies, the mutation rate of each strain and the proportion of each mutation type. "Poison?" controls the addition of an antibiotic that kills all nonresistant cells. Plots represent the total number and proportion of each strain over time.

expected to have a higher mutation rate and why. Then groups discussed data that suggested, contrary to most students' expectations, that *E. coli* have a relatively low per base pair rate of mutation (from Lynch, 2010). In these discussions, students raised many different possible explanations and questions about what factors and mechanisms could account for variable rates of mutation within and across species (e.g., genome size, ploidy, generation time, stability of environment, cost of DNA repair).

Additional discussion took place before beginning the experimental protocol. Students were asked to predict what they might see on each experimental plate (nutrient-rich LB vs. LB with antibiotic added). GTAs were instructed to ask students to explain the reasoning behind their predictions. Students often predicted that there would be more colonies of the fast-mutating strain on the antibiotic plates due to mutations that conferred resistance, but students were less sure what to expect on the nonselective LB plate. After the whole-class discussion of predictions, student groups conducted the same basic experimental procedure as in the original lab. The following week, students analyzed their plates, and GTAs again led a discussion about what the results might mean.

Guidance for how to lead lab discussions was embedded in the 3-hour laboratory prep session. The lab coordinator (J.G.) explained that the purpose of these discussions was to get students thinking rather than to come to a correct or complete explanation. In lab prep sessions, the coordinator modeled discussions by asking for GTAs to share ideas and respond not with evaluation or correction, but rather with requests for elaboration (e.g., "In what way are you thinking a large genome size could affect mutation rate?"). While we do not have direct evidence of how individual GTAs implemented these discussions, we have indirect evidence of their attempts from the debriefing discussions. During these discussions, GTAs described the ideas, some expected and some unexpected, raised by students in their lab sections.

While the experiment was essentially the same as in the original lab, we did introduce a new activity designed to complicate the experimental results: an agent-based computer simulation that allowed students to grow virtual strains of bacteria with different mutation rates. On the simulation interface, students could change the relative proportions of mutation effects (lethal, neutral, antibiotic resistance) and monitor the relative population sizes of faster- and slower-mutating strains over time (Figure 2). Students could also change the environment by adding or removing antibiotics. The activity, which lasted about 30 minutes at the end of the unit, asked lab groups to find conditions that favored the faster-mutating strain, conditions that favored the slower-mutating strain, as well as conditions in which the two strains could coexist.

A key feature of the simulation was that it made it possible for students to observe multiple possible outcomes, including some that students found surprising. In line with students' expectations, the strain with the higher rate of mutation could gain resistance mutations in an antibiotic environment and increase in frequency. Without antibiotics, the slower-mutating strain would quickly gain an advantage if lethal rates were above zero. Less obviously, the faster-mutating strain would often decline in frequency over time. Once it had fixed a beneficial mutation, the faster-mutating strain began to suffer effects of deleterious mutations (including back mutations). In addition, because of the inherent stochasticity of an agent-based model and the sensitivity of dynamics to parameter values, students could observe variation across runs (for more details on the role of the computer simulation in the design, see Gouvea *et al.*, 2022). While these patterns were possible for students to observe, GTAs reported that few students had sufficient time to notice these trends. Often, students spent most of the 30 minutes allotted to the simulation getting oriented to the interface. Students also seemed confused about the purpose of the simulation, and many seemed to view it as unrelated to the experiments they had conducted.

Another major change in this design iteration was in the guidelines and grading structure for the lab report assignment (see Appendix A in the Supplemental Material). Here, we explicitly stated that the purpose [of the lab report] is not to get to some particular "answer," but rather "to make some sense of the data and support your ideas with logic and evidence," and that reports would be graded based on "clear expression of ideas and evidence." The discussion section was described as a place to answer the question: What does the experiment tell you about some of the questions you raised in the Introduction?" While the emphasis was on experimental results, the discussion section guidelines also suggested the option of including simulation output. In addition, we extended the peer review process to include the introduction, methods, and results. Pairs of students swapped drafts of their lab reports for feedback outside class and had 15 minutes in class to discuss feedback with one another and time outside class to revise before submitting a final version.

Second Design Iteration. In the second design iteration, we preserved the small-group and whole-class discussion activities introduced in the first design iteration. The focus of this second iteration was to make changes to the lab activities (Table 1). Our aim was to increase the complexity of the data that students encountered by making it possible for them to encounter contradictions or ambiguous patterns. We did this in several ways. First, we removed the lactose condition from the first experiments so that students would compare only the rich medium (LB agar) and the antibiotic (LB + rif) conditions. Our purpose was to de-emphasize adaptive benefits and make more space for students to consider how mutation rates influence fitness in nonselective environments. We also increased the dilution factor for the LB agar plates so that instead of seeing an undifferentiated lawn, students could count the colonies and quantitatively compare the two strains.

Second, we added a "competition" experiment that students performed in the second week. In this experiment, each group incubated the fast- and slow-mutating strains together in a single culture tube of LB medium overnight. They then plated the mixed culture on LB agar and on LB + rif plates to estimate the relative proportion of each strain.[1] Each lab section conducted an independent replicate of the competition experiment, and the results varied by lab section. Some results suggested the mixed culture contained a larger proportion of the faster mutator. Other results suggested a larger proportion of the slower mutator. Still other groups found a similar proportion. Students were given access to data from all 12 lab sections. We intended the lack of a single pattern to create a need to consider the validity and meaning of the data.

Finally, the computer simulation activity was extended to an hour and included some time for groups to share interesting patterns with the GTA, who could project simulation output for the whole class to consider. In pre-lab prep, GTAs were instructed to draw students' attention specifically to any patterns that they found puzzling or contradictory, such as runs in which the slower-mutating strain eventually overtook the faster-mutating strain in a selective environment. To facilitate students noticing these patterns, changes were also made to the simulation interface. In the new interface, students could run two replicates side-by-side. For example, students could run a trial in a rich medium and a trial in a selective medium at the same time and compare the results. In addition to the proportion of lethal and resistance mutations, the proportion of "metabolic benefit" mutations were now manipulatable by students so that various hypotheses about how the relative proportions of different types of mutations and population fitness could be tested. The updated simulation also included plots that allowed students to visualize the proportion of the population that had acquired metabolic benefit or antibiotic resistance mutations, allowing them to make connections between population patterns and underlying patterns of mutation frequency.

Instructions for completing the lab report remained similar to the 2015 version. We continued the practice of peer review, but this time we had students share a draft of their discussion sections only, instructing students to attend to the reasoning their peers provided in this section.

Data Collection

All student lab reports were submitted digitally each year, and copies were stored in an online database.[2] We sampled from these reports in two ways. The first subsample consisted of all the lab reports from a single GTA who taught in all 3 years. This GTA was a graduate student in the biology department who was personally committed to teaching. In discussions, this GTA confirmed that their teaching in 2014 followed the template laid out in the lab manual: They began each lab with a short lecture and then helped guide students through the experimental protocol. They also indicated an intention to implement the redesigned labs with fidelity. This intention was supported by informal observations of this GTA's classroom in which we observed them eliciting students' ideas during discussions and asking for students' reasoning. This single GTA subsample included 24 reports from 2014, 19 from 2015, and 25 from 2016 for a total of 68. A second subsample was constructed by selecting a random set of 78 (26 per year) reports from among the remaining GTA sections. Summing these two samples, 146 reports were collected for analysis.

Scoring Argumentation in Lab Reports

We used a modified version of the structure of observed learning outcomes (SOLO) taxonomy (Biggs and Collis, 1982) to score students' engagement in argumentation in lab reports. The original SOLO taxonomy was developed to assess the quality of students' writing using criteria that could be applied across a range of domains, including geography, history, and English. Across domains, Biggs and Collis (1982) define quality in terms of how students use and integrate evidence as well as the extent to which the writing includes evidence that students' interpretations are their *own* (p. 54, emphasis in original).[3]

[1]To do this, students estimated the proportion of cells in the mixed culture that were resistant to rifampicin. They then compared this proportion to the proportion of resistant cells for each strain independently, which had been estimated the week prior. A larger proportion of resistant cells would suggest a larger proportion of the faster-mutating strain in the mixed culture and vice versa.

[2]This data collection and analysis was approved by Tufts University Institutional Review Board (IRB no. 1904014).

[3]Biggs and Collis originally based their scheme on Piagetian levels of cognitive development, expecting that lower-level responses corresponded to lower levels of individual development. However, in applying their scheme, they found that "stage theory did not hold"— individual students were capable of writing at different levels (1982, p. 21). Biggs and Collis revised their interpretation of the scheme from one that assessed cognitive development of the *student* to one that assessed the quality of the *task* (1982, p. 22).

While Biggs and Collis (1982) never use the term "framing," their scheme is organized around capturing features of writing that indicate how students are understanding and approaching the task: Are they writing to demonstrate knowledge or fulfill requirements, or are they writing to express and defend their own ideas? Thus, we felt it was appropriate to use this scheme to characterize engagement in scientific argumentation that was more rote or more authentic.

Biggs and Collis identified three dimensions as useful for scoring responses. In the following sections, we describe how we adapted these dimensions to fit our context. It should be noted that, while these dimensions are useful for determining

TABLE 2. Description of dimensions that comprise each level with example student text and justification provided by coder.

	Level 1	Level 2	Level 3	Level 4
Argument structure	**Tautological/simple**	**Additive/one-sided**	**Relational/two-sided**	**Compound/conditional**
Example text *[coder justification]*	The results above conform to the hypothesis that as the efficiency of DNA repair mechanisms increases the mutation rate decreases. *[Known difference in DNA repair used to explain known difference in mutation rate]*	Combining the data from Part I and II, it can be deduced that the E938 strain's status as mut– [high mutating] allows it to better survive. *[Makes claim about success of high mutator only]*	Overall, since there are so many things that affect the reproduction of bacteria, such as when a mutation occurs, what kind of mutation occurs, and human error, it is difficult to use our limited data to determine which strain would be more suited for survival. However, the general trends of our data suggest that having a high mutation rate is beneficial to survival. *[Makes claim about success of high mutator, but only after relating to other possible claims]*	It might seem like a good idea to argue against DNA repair. However, due to the staggeringly small percentage of the population that develops beneficial mutations, it does not make evolutionary sense to argue this point. *[Considers advantages of high and low mutators and reconciles using evolutionary argument]*
Scope of knowledge/ evidence	**Relies on given information and/or unspecified results**	**Uses subset of available data as evidence to support inference**	**Relates data that show different patterns**	**Relates data and brings up outside information or hypotheticals/ thought experiments**
Example text *[coder justification]*	Our results showed that E938 (Mut–) was more likely to mutate than E939 (Mut+). *[Mentions "results" without explanation or elaboration. Facts about mutation rates given.]*	In Part I and II E938 was able to grow more colonies on plates containing an antibiotic and in Part II E938 was able to adapt to using lactose as a food source. *[Combines two parts of data to support claim with some explanation, ignores agar data]*	The lower mutation rate … does not allow it to mutate enough to generate a significant number of bacteria that could digest lactose, … On another note, both bacterial strains survived roughly equally as well in the environment with just LB Agar. *[Considers data from all environments]*	Together these data provide support for the hypothesis that an increased mutation rate was beneficial to bacterial survival. …. While more mutations appear to be an advantage, the relationship between mutation rates and repair mechanisms remains intriguing. If these results hold true in nature, then why is DNA repair so pervasive? Statistically, with more mutations comes a higher chance for an organism to mutate in a non-beneficial manner as well. *[Considers consistency between data and patterns in nature including statistical hypothetical argument]*

(*Continued*)

TABLE 2. Continued

Consistency and closure	Closed without rationale, ignores inconsistencies	Closed with some rationale, ignores inconsistencies	Closed, addresses inconsistencies	Appropriately qualified or open-ended
Example text *[coder justification]*	The results of both parts of the experiment conformed to what was expected. *[Concludes with certainty due to match to expectations]*	This observed pattern was consistent with what was seen in the lab class [as] a whole, indicating that the results of this study are reproducible. *[Uses reproducibility to justify certainty, but does not address other inconsistencies]*	Overall more mutations can be beneficial for a species as a whole given that there is a higher possibility that the species will be able to adapt to changing environmental factors; however given more mutations there is a higher possibility that individuals will not be able to survive in a given environment. Essentially, a higher rate of mutation adds to the genetic diversity in a species, which has been proven to be an important factor in a species ability to survive *[Addresses possibility that mutations are not beneficial, yet still concludes with general claim about advantage]*	While this experiment successfully displays the advantages of a high mutation rate in an unfamiliar environment, the similar lawns of bacteria that appeared on the LB plates limit our ability to determine the effects of mutation in a non-hostile or unfamiliar environment. *[Explores possibility of different outcomes in different contexts]*

an overall score, they are not intended to be mutually exclusive. Examples of how we scored lab reports using these dimensions are provided in Table 2,[4] and complete examples of reports for each level with text used to assign scores are provided in Appendix B in the Supplemental Material.

Argument Structure.[5] This dimension captured the complexity of the argument in terms of the number of claims and the extent to which claims were related to one another (cf. Schwarz *et al.*, 2003; Osborne *et al.*, 2016). We also considered whether the claim itself could have originated in information that was given in the lab manual or lecture or if it was inferred from data patterns. For example, readings across all three versions of the course described that a dysfunction in the DNA repair system was responsible for the higher rate of mutation in one strain. Level 1 reports claimed to demonstrate this given information, arguing that bacteria with a mutation in their DNA repair system would mutate more—a claim that is tautologically true and cannot be inferred from the experimental data.

Level 2 arguments also made a single claim: A higher rate of mutation is advantageous. This claim was suggested by the lab manual in 2014, but not in other years. This claim could also be inferred from the data showing more survival of the faster-mutating strain in an antibiotic environment. Level 2 claims were considered "one-sided," because they did not address the potential fitness costs or context specificity of advantage of an increased mutation rate.

Level 3 arguments were "two-sided," considering both that faster mutation could be adaptive and that more mutations could potentially harm the population. While level 3 arguments raised both claims, they often did not fully reconcile them, choosing instead to align with one or the other.

At level 4, alternative claims were raised and integrated into compound arguments that provided conditions explaining when different claims might be true. For example, many level 4 reports described advantage as environment dependent, dynamic, or sensitive to specific parameters, such as the proportion of deleterious or beneficial mutations, population size, the rate of back mutations, or the relative strength of selection.

Scope of Evidence.[6] This dimension characterized the sources of evidence and how they were used to support claims. At level 1, either no reference was made to evidence or else "results" were taken as evidence without explanation. In such reports, results were positioned to self-evidently "show" conclusions. At level 2, reports used specific evidence, but only a subset of the available evidence. For example, level 2 reports primarily used data showing the increased survival of mutants on antibiotic media and/or an increased ability to digest lactose as evidence of fitness advantage of the fast mutator. Reports in this category

[4]The original SOLO taxonomy had five levels. Our adaptation has four, because we had no instances of what Biggs and Collis called "unistructural" responses that relied on a single piece of evidence. This is likely because all versions of the lab involved more than one set of data.

[5]Biggs and Collis originally called this dimension "relating operations" to describe the logical structure of students' claims and the reasoning that accompanied those claims. Because of the combined focus on claims and reasoning, we renamed this dimension "argument structure."

[6]Biggs and Collins originally named this category "capacity" to describe the amount of "working memory" required to recall information needed to write responses at different levels (1982, p. 26). In keeping with Biggs and Collis's shift from interpreting their scheme as about describing tasks rather than inferring cognition, we have renamed this dimension to foreground the evidence present in responses.

largely ignored growth in the standard LB media, either not mentioning it at all or describing growth in standard plates as a "control." In contrast, level 3 reports used both data from selective and nonselective media as evidence.

Level 4 reports increased the scope of evidence to include knowledge from beyond the pre-lab readings and the experimental data. These reports often integrated theoretical knowledge in the form of thought experiments or analogies to complicate interpretations of data. For example, some reports included the fact that DNA repair exists and is pervasive across taxa as a form of evidence that challenges interpretations of mutations as unilaterally advantageous. Other reports included more idiosyncratic ideas. For example, one student constructed a metaphor based on the "X-men" comics series to explain how the benefits of mutation are context dependent. In the comics, mutation gives the X-men superpowers. However, in one storyline, robot sentinels are designed to be able to specifically detect and target these mutants. In other words, the ability to mutate is both the source of their power and a threat to their existence, which the student used to explain that "Whether a mutation prove [*sic*] to be beneficial or detrimental, and whether a higher mutation rate is advantageous or disadvantageous is situational." This example illustrates how the use of evidence in level 4 reports was more varied than in the other levels and sometimes contained idiosyncratic ideas.

Consistency and Closure.[7] This dimension captured the extent to which students' conclusions acknowledged uncertainty by either leaving conclusions "open" or, in contrast, "closing" on a single conclusion. When students feel pressured to present conclusions that align with the scientific canon, they may "prematurely close" their arguments (Engle and Conant, 2002; Ford, 2012). Researchers have used linguistic markers of ongoing uncertainty, such as students' use of appropriate qualifiers, conditionals, or hedges to signal appropriate uncertainty in arguments (Lee *et al.*, 2017; Chen *et al.*, 2019). We used such linguistic indicators of certainty to score this dimension.

Level 1 reports contained conclusions that closed on expected claims with certainty. These reports included language such as "showed," "proved," or "confirmed." Level 2 reports also settled on conclusions, but whereas level 1 reports did not attempt to justify this settling, level 2 reports offered some backing for their certainty. For example, level 2 reports sometimes used consistency across the two experiments or among groups in a lab section to strengthen their conclusions.

In a break from the certainty presented at levels 1 and 2, level 3 reports acknowledged tensions in interpretation, often raising the idea that mutations can be advantageous or deleterious. However, level 3 reports did not fully reconcile this tension. Instead, they often simply chose a side on which to conclude. For example, after discussing data in terms of benefits and harmful effects, one level 3 report concluded: "From an evolutionary biological perspective, mutations have a very big and important role in developing new species that will outcompete the weaker species that fail to adapt to their surroundings." Rather than allow the ambiguity in the data to stand, this report brushes it away by making reference to the importance of adaptation.

Only at level 4 were uncertainties more directly expressed as limitations (e.g., "it would be a mistake to conclude that mutations are always beneficial"). Level 4 reports often included hedging language (e.g., "may be," "could be considered"). Some level 4 reports also raised new questions or suggested the need for future work (e.g., "further studies should be done to explore the long-term survival of bacteria with higher mutation rates—particularly considering the rates and consequences of negative, as well as positive, mutations"). Proposals for future work were also sometimes part of reports at lower levels, but these ideas never arose from uncertainties. Instead, level 1 and 2 reports often proposed future work that replicated or made minor changes to the experiments already conducted (e.g., "it would be interesting to alter the environment and see how *E. coli* with no DNA repair could grow in heat or cold").

Coding Process

Before coding, we excised the discussion sections from sampled reports, removed identifying information, and ordered them randomly in a spreadsheet. This process blinded coders to the implementation year of individual responses. It should be noted, however, that it was possible in some cases for coders to tell what year a report likely came from, because some of the details of the lab activities changed over the years. For example, a report that mentioned simulation output could not have been written in 2014 before the simulation activity was introduced. To reduce the influence of potential bias, coders identified specific words or phrases used to assign a level in each of the three dimensions, bolding or underlining the parts of the text they used to make their judgments and making notes to justify their choices and to indicate their degree of certainty (see examples in Appendix B in the Supplemental Material).

While different parts of the text (claims, evidence, language related to (un)certainty) informed scoring in each dimension, the dimensions themselves are not intended to be independent of one another. In Biggs and Collis's (1982) original scheme, each of the three dimensions of the taxonomy were expected to correlate and to be considered holistically to assign an overall best fit to a single level. In our data, dimension scores were often aligned, making it relatively easy to assign a writing sample to a single level. However, Biggs and Collis also created "transitional" levels to capture responses with split dimensional scores that spanned levels (1982, p. 29). We followed both practices in our coding. Coders first scored for each dimension and then assigned an overall score. When dimension scoring was split across levels, we discussed the examples and chose an appropriate transitional score. Examples of reports coded at each level, including transitional levels, are provided in Appendix B in the Supplemental Material.

Four researchers (the four authors) participated in the coding process. Each report was scored by at least two coders independently. Any discrepancies were discussed and resolved between the two coders. Particularly challenging examples were brought to the larger group for more discussion. Because the SOLO categories are ordinal, we calculated interrater reliability of initial preconsensus coding of the full data set using Cronbach's kappa with linear weighting (Cohen, 1968). The

7Our use of this dimension closely matches the original use by Biggs and Collis (1982, p. 27).

weighted kappa was 0.74 (SE = 0.03), corresponding to a "good" level of agreement.

Statistical Analysis

To compare the distribution of responses in each level across years, we used a Kruskal-Wallis test, a nonparametric rank test (analogous to analysis of variance). We then used the Dunn test to conduct post hoc pairwise comparisons between all years.

Reports sampled from a single GTA ($N = 68$) were compared with reports sampled randomly from across all GTA sections ($N = 78$) using a Kruskal-Wallis test. Because no statistical difference was detected between these subsamples ($Z = 0.64$, $p = 0.52$), we combined the subsets into a single data set for the analysis.

Characterizing References to and Use of Computer Simulation Output in Lab Reports

The agent-based computer simulation was a unique aspect of the curriculum included in 2015 and expanded in 2016. Including and discussing patterns from the computer simulation in lab reports was recommended, but optional in both 2015 and 2016. To better understand the role of this activity in students' arguments, we counted the number of sampled reports in each of the redesign years that included explicit mention of simulation output.

We then categorized how students used the simulation output in their arguments. Some *aligned* the simulation output with the experimental data. Below is an example of a report that aligned the simulation output and experimental data to claim that mutations that confer antibiotic resistance are beneficial:

> The NetLogo simulations in lab confer with these results. In the simulation, adding [antibiotic] to the environment resulted in more yellow bacteria, the [antibiotic-resistant] mutant of the red strain. The blue strain died off because the red strain's high mutation rate increased its resistance to the [antibiotic]. (2015, level 2.5)

Other reports used simulation output to describe a *range of possible outcomes* and explain this variation in terms of dependence on parameter values (relative proportions of deleterious, beneficial, and neutral mutations, or rate of back mutation), population dynamics (fixation of beneficial mutations or accumulation of deleterious mutations over time), or environmental conditions (e.g., presence or absence of antibiotics). The following example uses the simulation to expand from the specific claim—that a faster-mutating strain of *E. coli* can have an advantage in an antibiotic environment—to make a more general claim about the situation specificity of advantage.

> The experiment was only an example that can conclude the benefits of a higher mutation rate to specifically E. coli [*sic*] specifically under specific environments. Through NetLogo simulation, we were able to expand our knowledge to make a more general claim: a specie's [*sic*] survival depends on a combination of its mutation rate and the effects of mutations in the given environment. The simulation can be configures [*sic*] such that a higher mutation rate will lead [to] complete dominance, total extinction, or every variations in between. The same applies to a low mutation rate. This means that the evolutionary advantage of mutations and mutation rates are not clear-cut. Whether a mutation prove[s] to be beneficial or detrimental, and whether a higher mutation rate is advantageous or disadvantageous is situational; it entirely depends on the environment that one lives in. (2016, level 4)

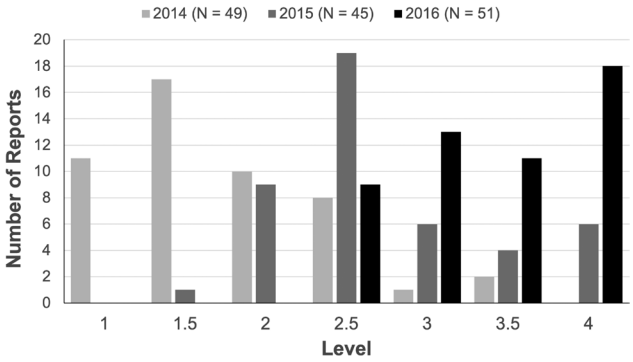

FIGURE 3. Lab report scores by year. Reports from 2014 in light gray; 2015 in medium gray; 2016 in black.

The above examples illustrate the two categories we used to classify how students used simulation output: aligning with the experiment or complicating or extending claims by discussing multiple possible outcomes.

RESULTS

In this section, we present the results of our coding and scoring of the 146 laboratory reports. We first show that the scores we assigned to lab reports increased each year. We then present excerpts from two example reports (a low-scoring report from 2014 and a higher-scoring report from 2016) to illustrate how these scores reflect students' engagement in argumentation. Finally, we describe patterns in how students used the simulation output in their arguments.

Lab Report Scores Increased with Each Design Iteration

Figure 3 shows the distribution of lab report scores by year. Over the two design iterations, the distribution shifted toward higher levels in the coding scheme, indicating more authentic (or less rote) engagement in argumentation ($\chi^2 = 79.7$; $df = 2$; $p < 0.0001$. In the original labs (2014), the median and mode scores were level 1.5; in the first design iteration (2015), the median and mode shifted to level 2.5, and finally, in the second design iteration (2016), the median shifted to level 3.5, while the mode increased to level 4. Pairwise comparisons indicate that each shift is significant: between 2014 and 2015 ($Z = 4.9$; $p < 0.0001$); between 2015 and 2016 ($Z = 3.7$; $p = 0.0006$).

An Illustrative Comparison of Lab Reports

To illustrate the relationship between the numerical trends and engagement in argumentation, we share excerpts from two example reports. The first is from the original lab

course (2014) and was scored at level 1.5, the most common score assigned to reports from that year. The second was sampled from the second design iteration (2016) and was scored at level 4, the most common score in that year. Excerpts have been edited for length, but the full discussion sections of these two examples, as well as example reports for all scores, are available in Appendix B in the Supplemental Material.

Example of a Level 1.5 Report from 2014. This level 1.5 report begins with the following excerpt:

Line	Text
1	Our hypothesis was that [fast-mutating strain] E938, which is Mut-, would show increased mutation over [the slow-mutating] strain E939.
2	This hypothesis was supported by both our individual and class data.
3	The increased mutation rate and survival of E938 can be attributed to the faulty DNA repair of the Mut– strain.
4	Increased mutation rate allows more bacteria of this strain to gain the traits necessary for survival on the rifampicin plates.
5	E938 also adapts better for the use of lactose as an energy source on the MacConkey agar.

This excerpt begins with claims that reiterate ideas provided in the lab manual. The claim that bacteria with faulty DNA repair will have an increased rate of mutation (line 1) restates a given fact about the strains from the lab manual ("Mut– develops mutations at a faster rate than Mut+"). In elaborating on this claim, the student also suggests that a faster-mutating strain has some advantage, because it "gains the traits necessary for survival" and "adapts better" (lines 4 and 5). These statements are also tied to the original lab manual, which states that "a mutation can cause an increase in fitness."

While these claims are not incorrect, we consider them to be evidence of rote engagement in argumentation for two reasons. First, while it is sensible to claim that faulty DNA repair increases mutation rates, the data that students collected in the lab do not and cannot support that link, because the students did not collect any data on DNA repair. The only link between repair status and the observed mutation frequency is the given knowledge that the two strains differ, making the claim that faulty DNA repair increases mutation rate tautological. The claim that mutations are beneficial is also sensible, but given that mutations can range from beneficial to neutral to deleterious, it is a relatively narrow claim. From a biological perspective, these claims present only a limited set of ideas about the phenomenon of mutation rate. The strong alignment between the claims and the information provided in the lab manual make it unlikely that these ideas represent the student's independent thinking about the meaning of the data. Instead, the inclusion of these claims appears intended to communicate the conclusion suggested by the curriculum.

The next excerpt illustrates how we evaluated the certainty of argument in this same report:

Line	Text
1	Some uncertainty was present in the class data for the [fast-mutating] E938 strain.
2	The standard deviations were high (30% and 40% respectively).
3	This may be attributable to problems while performing the serial dilution.
4	Groups may not have allowed for the proper dispersion of the bacteria in each solution before performing further dilutions.
5	This may result in initially lower concentrations of bacteria on some plates.
6	This can easily be remedied by allowing proper dispersion of the bacteria in each dilution step.
7	Despite the high standard deviation for the class E938 data, the experiment supported our hypothesis ...
8	... that E938 would show increased mutation due to the lack of a functional DNA repair mechanism.

In this excerpt, the student identifies potential threats to the validity of the data, while at the same time upholding the certainty of the conclusion. Specifically, the student first includes information about the high standard deviations of colony counts (line 2) and links this to a possible source of human error in conducting dilutions (lines 3 and 4). Then these issues are dismissed, and the conclusion is restated without qualification (lines 7 and 8). No further consideration is given to the possible biological meaning of the variation or how that variation might relate to the mutation rate. Again, it seems that the issues and potential sources of error are included for the sake of meeting the instructor's expectation about lab report components rather than because they are consequential to the overall argument being presented.

Together, these two excerpts illustrate common features of lab reports that spanned levels 1 and 2: Claims are often narrowly aligned with the background information provided in lab manual; evidence is included as supporting these claims with a high degree of certainty; and concerns about the quality of the data are, if included, dismissed. These features suggest a framing of writing a lab report as an activity for the purpose of "doing what is expected."

Example of a Level 4 Report from 2016. This level 4 report does not begin with a statement of the main claims and instead includes five different claims distributed throughout. This claim structure was more common in reports that were scored at levels 3 and 4. We present here excerpts of the claims in the order that they appear in the report:

1. "Our results somewhat supported the hypothesis, in that the mutant [faster-mutating] E938 strain had a much higher survival rate on the rifampicin plates than the wild-type [slower-mutating] E939 strain."
2. "The strains were possibly able to co-exist with one another because, although the strain with the higher mutation rate has a higher resistance to rifampicin, the wild-type strain is more stable altogether due to its functioning RNA polymerase and 'proofreading' system. Therefore, it doesn't experience as many deleterious mutations."

3. "There's a chance that the strain with the higher mutation rate would eventually die out due to harmful and at times fatal mutations."
4. "It's important to consider the metabolic cost such an extreme mutation rate has on bacteria. Although the mutant E938 could survive better in an environment with rifampicin, it has to compensate for this adaptation in other ways, meaning that its ecological range is most likely reduced."
5. "That relates to why organisms have such varying mutation rates, especially bacteria in relation to humans and plants. Billions of E. coli [*sic*] bacteria are produced in our intestines every day—along with millions of mutations within their populations. Since bacteria have such a short lifespan and rapid reproduction rate compared with plants and humans, their mutation rates are much more significant."

The report begins with the expected idea that a faster-mutating strain will survive better in an antibiotic environment (claim 1). However, notice that the results are introduced as "somewhat" supporting this hypothesis. The next two claims elaborate on why the survival data only somewhat support the hypothesis. Claims 2 and 3 introduce the idea that avoiding deleterious mutations is important and explains why a strain with a lower mutation rate would experience fewer deleterious mutations than a strain with a higher mutation rate. Claim 3 also includes a consideration of timescale, specifically attending to the possibility that a fast-mutating strain might "eventually die out." Claim 4 integrates different parts of the argument to articulate a trade-off whereby strains that mutate faster trade adaptive advantage for costs that may reduce the strain's overall success ("ecological range"). Finally, claim 5 considers phenomena beyond the experiment to account for differing rates of mutation across taxa, though it is not entirely clear what the student's intended meaning is in this last section.[8]

The argument structure of this report compared with that of the first example report offers a more complex, more complete discussion of the relationship between mutation rate and fitness. In addition, the claims of this report go beyond what was provided in the lab readings. And the fact that the final idea is somewhat difficult to interpret may constitute evidence that this final idea is the student's original thought.

The next excerpts come from the final paragraph of this (level 4) lab report in which the student engages with the quality of the data. Like the level 1.5 example, the student begins by identifying possible sources of error: "For example, plating the petri dishes required the serial dilution of each of the strains multiple times and in very small dosages, which left a lot of room for human error." One difference from the level 1.5 report, though, is that this error is not dismissed and instead the student suggests a need to "re-do the experiment."

A more telling difference between the two reports is that this report ends with an idea for a future study that is inspired by unresolved questions from the previous analysis:

> A way to further our study is to continue growing the co-culture for a longer amount of time than just a week to see what eventually happens, such as whether or not the higher mutation rate is overall [more] harmful to the mutant population than beneficial.

Up to this point, the student has made the case that both high and low rates of mutation have potential benefits (claims 1 and 2). They have also suggested that perhaps a higher rate of mutation could be detrimental over longer periods of time (claim 3)—a claim they derived from observations in the computer simulation. By proposing an experiment that is extended in time (more than 1 week), this student is proposing a next step that addresses a limitation of the current experiment as well as a gap revealed by their own argument.

Overall, this example report, and others at level 3 and above, include more evidence of students' own thinking about how claims are related as well as how claims might relate to the biological world beyond the lab. In addition, in these reports, students' use of common features of a lab report, such as reporting on errors or proposing future experiments, is integrated with the rest of the report. These features suggest a framing of the lab report as an activity that involves creative and critical thinking about scientific interpretations and ideas.

Use of Computer Simulation Output in Arguments

Overall, of the 96 reports sampled in 2015 and 2016, 27 (28%) made explicit use of the simulation output in their lab reports. Fewer reports referenced the simulation in 2015 (8 of 45, or 18%) than in 2016 (19 of 51, or 37%). How simulation output was used also differed by implementation year. In 2015, five reports used simulation output to align with initial experiments, and three used output to complicate and extend initial claims. In 2016, only three used the simulation to align and 16 used the output to complicate their claims by describing and explaining the meaning of multiple possibilities.

In both years, how students used the simulation was related to their overall argument score but did not fully determine it. Using the simulation to confirm that mutations are beneficial tended to occur in lower-scoring reports (below level 3), whereas using the simulation to describe context-dependent alternatives was a strategy present in high-scoring reports (levels 3 and 4). Yet reports that used the simulation to align with experimental claims could also score higher if they used other evidence (from experiments or outside knowledge) to make more complex arguments. For example, after initially using the simulation to confirm the adaptive benefit of mutation, a student argued against this claim later in their report adding, "Taking a look at the broader results, however, suggests that in reality, a higher mutation rate could be a detriment" (2016, level 4). Similarly, describing the multiple outcomes of the simulation did not automatically elevate the overall score if that information was not integrated into the argument. For example, one student explained how changes in parameters and environments led to different outcomes in the simulation but dismissed these patterns as "not related" to the experimental data (2015, level 2.5).

DISCUSSION

In this design experiment, we saw shifts in how students engaged in argumentation in their biology lab reports. The

[8]One possibility is that the student is suggesting that, because bacteria have faster generation times, they suffer more from deleterious mutations than organisms with longer life spans (e.g., humans). The student may be using this idea to explain why bacteria have a lower rate of mutation per generation than humans.

target of our design efforts was the rhetorical context—the material and social surroundings that authors (both scientists and students) use to frame the purpose and form of their writing. Differences in students' writing across the three versions of the laboratory course provide insights into how specific designed features of the curriculum and instruction may have functioned to change how students framed and implemented the task of writing. We next discuss how these design features function in each design iteration.

Original Lab Design: Meeting Instructor Expectations and Keeping Uncertainty Low

The high proportion of level 1 and 2 arguments in discussion sections in the original lab course suggests that most students were framing writing as an activity of demonstrating predetermined conclusions and following directions. For many students, these framings may have been familiar, as traditional K–12 lab science contexts often convey to students that "they must somehow generate, copy, or paraphrase the knowledge claim that is desired by the teacher" (Keys, 1999, p. 125). The expectation that science labs are about "confirming concepts" appears to persist at the undergraduate level as well (Hu *et al.*, 2017; Smith *et al.*, 2020).

Features of the design of the original lab course likely reinforced this framing. The lab manual, for example, presented canonical ideas and background facts that most students seemed to recognize as the desired conclusions. Indeed, most reports were organized around one of two ideas presented in the lab manual: 1) that the difference in mutation rate between the two *E. coli* strains could be explained by differences in DNA repair or 2) that the data showed that mutations are beneficial for adaptation.

The importance of "following directions" was reinforced in the structure of instruction and assessment. That GTAs delivered lectures and graded pre-lab quizzes may have positioned them as authority figures who knew both how to correctly implement procedures and what the data should ultimately mean. In introducing the lab report assignment, GTAs reviewed the multi-item rubric they would be using to grade the reports. Many of the rubric items asked for specific formatting (e.g., the title must include a dependent and independent variable), and overall, the instructions emphasized the need to follow directions carefully. This assessment structure can activate a "checklist" approach to writing (Tang *et al.*, 2015) and can explain why lower-level reports often included rhetorical features such as lists of sources of error or ideas for future studies that seemed tacked on and disconnected from the rest of the argument.

In terms of the rhetorical context, the original lab curriculum kept uncertainty low by explicitly suggesting what the data were supposed to show. At the same time, instructors were positioned as an audience concerned with checking that students met the expectations laid out in the detailed rubric. In this context, it makes sense that students would produce simple, even tautological arguments and use data to unambiguously support the conclusions they knew they were meant to support.

First Year of Redesign: An Audience That Values Student Thinking

We saw an increase in higher-scoring lab reports from the original year to this year. Most reports in 2015 were scored at level 2.5, which meant that, while many students were still focused on the one-sided claim that mutations are beneficial, they were beginning to complicate this argument. For example, a level 2.5 report might raise concerns such as the following: "[This experiment] does not really answer the question of how mutants would fare in normal environments" (Appendix B in the Supplemental Material). In addition, the proportion of level 3 and 4 reports increased from the original year to this year, further indicating that students began to engage in more creative and critical thinking and writing with the redesign.

One design change that can explain these results was our removal of background information from the lab manual that hinted at the expected claims. Research by Petritis and colleagues (2021) has also shown how consequential background information can be for how students frame labs. In their study, which took place in an introductory chemistry laboratory course, one group of students received information about the molecular structure of reactants before collecting experimental data. Another group of students was not given this information and instead had to observe the reactions and make inferences based on their observations. Lab reports written by students who were given the chemical structures ahead of time contained arguments that were less explicit, less complete, and less well integrated. Petritis *et al.* (2021) argue that the information about the structures cued up a framing that was focused on *verification*. In contrast, students who had not been given this information tended to interpret results as uncertain, creating a reason to consider more than one interpretation and, ultimately, yielding more comprehensive arguments that integrated evidence more carefully and completely.

To further communicate an emphasis on data interpretation over verification, we introduced the lab unit with the following question: Is it better to mutate a lot or a little (and under what conditions)? The phrasing of the question itself suggests the possibility of multiple answers. While it was still possible to answer the question with a one-sided answer—to argue for example that fast mutators should have an adaptive advantage—more students in this design iteration expanded the range of evidence they attended to, and more students constructed complex claims that explained how this advantage was part of a trade-off or limited to certain environments.

At the same time, in this design iteration, we shifted the instructor roles and assessment structures to increase the emphasis on students' thinking and decrease the emphasis on compliance with specific formatting guidelines. This may have helped position students as authors of ideas and GTAs as an audience interested in students' reasoning. Prior work has shown that how students understand their relationship with a scientific audience can influence their writing. For example, O'Neill (2001) showed that the nature of interactions between high school students and scientist mentors impacted how students wrote research reports. When the interactions involved more questioning, students' writing contained markers of critical argumentation, including hedging language, discussions of flaws or limits on conclusions, and considerations of alternative hypotheses. O'Neill (2001) argued that these pairings worked best when science mentors acted as a "responsive and critical audience" who could model how and when to be skeptical. Our assessment structured mirrored this emphasis on students' thinking by replacing the multi-item rubric with instructions

that asked students to "support your ideas with reasoning" and by having GTAs demonstrate attention to students' ideas in discussions.

Nevertheless, in 2015, a large proportion of reports still contained simple confirmatory arguments relative to those written in 2016. We attribute this pattern to differences in the complexity of the data we asked students to think about in these two design iterations. In 2015, the main pattern students could observe in the experiment was that a faster-mutating strain of bacteria was more likely to adapt to a novel environment (e.g., with antibiotics). While some students in 2015 used data from the nonselective (LB agar) plate to examine the limits of the advantage, many students that year still interpreted the patterns as straightforwardly confirming their initial expectations. Moreover, the brevity of the simulation activity in 2015 and the fact that it was presented at the very end of the unit seemed to communicate to students that the simulation was either intended to support their expectations or not relevant to their reports.

Second Design Iteration: Uncertainty Motivates More Complex Arguments

In the second design iteration, we focused our design efforts on introducing more uncertainty into the lab unit (Manz, 2015b). Specifically, we introduced an additional experiment, coculturing strains, which evoked multiple plausible predictions and did not straightforwardly rule out any of these interpretations. Because the data both within and across lab sections did not show a clear single pattern, the alternative explanations were kept "in play" in students' reports. As illustrated in the level 4 report (Appendix B in the Supplemental Material), students saw results that "somewhat supported the hypothesis" and then went on to discuss possible explanations that could account for multiple possible outcomes.

The simulation also functioned to create disciplinary problems for students to engage with in their reports. For some, the long-term patterns in the simulation contradicted expectations about the advantage of higher mutation rates by showing outcomes that favored strains with lower rates of mutation. Such discrepancies created a need for students to explain both outcomes (Blikstein *et al.*, 2016). The simulation also supported students in making more carefully qualified claims by allowing them to observe both the stochasticity of the system and the parameter dependence of outcomes (Gouvea and Wagh, 2018; Gouvea *et al.*, 2022).

Because the design of instruction and assessments remained similar between 2015 and 2016, we attribute the shifts in argumentation to the changing context, which now presented uncertainties, ambiguities, and contradictions for students to grapple with in their lab reports. A similar design approach was taken by Hester and colleagues (2018), who intentionally designed biology labs around phenomena for which it is possible (and plausible) to propose more than one explanatory model. Different student groups proposed different models, creating the potential for students to use argumentation to resolve discrepancies. While Hester *et al.* (2018) did not conduct a systematic analysis of students' writing, examples of lab reports provided in their paper show evidence of engaged argumentation: Students used lab reports to propose and provide support for their own models, rule out alternative models, describe the limitations of their models, and identify gaps or contradictions that remained unresolved.

Implications for the Design of Learning Environments to Support Scientific Practices

That students wrote more complex arguments in response to more complex data patterns may seem unsurprising. Yet it is important to remember that often the simplicity of students' arguments is attributed to a lack of ability on the part of students. In this experiment, we have shown that introductory-level students rose to the challenge of the rhetorical context without explicit instruction telling them that a good scientific argument considers multiple claims, integrates a range of data, including prior knowledge, and makes careful conclusions that do not overgeneralize or prematurely settle on conclusions. Instead, students did these things on their own when presented with complex data and when encouraged to present their own thinking rather than conform to the expectations suggested by instructors or curriculum. More generally, we see these data as supporting the core idea that people (both students and scientists) write in contexts, and those contexts, not simply their knowledge or skills, function to motivate and shape what and how they write.

This design orientation has broader implications for the design of learning environments that seek to support engagement in scientific practices, a central goal of biology education (American Association for the Advancement of Science, 2011) and science education broadly (National Research Council, 2012; Ford, 2015). In our section *The Role of Context in Scientific Argumentation*, we contrasted two approaches to teaching scientific practice. One has been to identify and scaffold the component knowledge and skills that are needed for competence with practices. Argumentation, for example, requires teaching students the knowledge and skills they need to support claims with evidence, to consider the quality of evidence, and to attend to and rebut counterarguments. Some scholars have emphasized the need to explicitly introduce and scaffold these specific skills (Kelly and Bazerman, 2003; Osborne *et al.*, 2004; Moon *et al.*, 2017).

Others have argued instead for the importance of designing contexts that create the conditions from which scientific practices are likely to *emerge* (Engle and Conant, 2002; Ford, 2005, 2012; Manz, 2012, 2015a,b). This approach shifts the scope of the problem from students and their individual capacities to a consideration of how individual behavior emerges in interaction with context. This alternative does not require removal of all support and all scaffolding. Our lab report instructions, for example, did contain specific guidelines and suggestions. This guidance asked students to focus on making sense of the data they collected and explaining their thinking. But in our study, guidance alone was not sufficient to promote more authentic engagement in argumentation; the lab activities also had to present students with data that needed to be interpreted. While it is common to think of uncertainty as a barrier or a challenge that students must overcome, we suggest that uncertainty can itself function as a kind of support for scientific activity. Uncertainty is a feature of science that motivates scientific work. In this sense, encountering uncertainty is an opportunity that is present in scientific communities but too often absent in instructional contexts. Thus, we are not arguing for a reduction of

support, but for an expanded understanding of the nature of support beyond scaffolding specific skills and structures to encompass purpose and function (cf. Manz, 2015a). The design challenge is to construct systems of activity in which scientific practices such as argumentation are perceived by students as sensible and valuable things to do.

A key consideration in the design of learning environments, then, is how various practices—including but not limited to writing—can and do function in scientific communities beyond the classroom. For example, experiments are useful for isolating and comparing effects, modeling is useful for articulating relationships and mechanisms, and statistics are useful when there is uncertainty over whether an observed effect represents a meaningful difference. When the problems that these practices have been developed to address are not present, neither students nor scientists have an intrinsic motivation to engage in them.

Crucially, the argument for designing contexts from which practices can emerge hinges on the assumption that students already have some productive knowledge and skills that can be applied to their own scientific practice. We believe this was true of the students in our study, who were sampled from a population of first- and second-year students at a private institution. While we cannot make general claims about what knowledge and skills students may bring to other institutional settings, we do argue that it is likely to be the case that many undergraduate contexts currently underestimate students' capacities. Indeed, research in K–12 contexts provides many examples of young learners' expanded capacities to engage in argumentation when contexts offer an interested and supportive audience and problems to think about (Metz, 2004; Berland and Reiser, 2011; Ford, 2012; Ryu and Sandoval, 2012). As Manz (2015a) argues, the mechanism at play in these situations is not direct instruction or explicit scaffolding, but rather the embedding of argumentation within the inherently uncertain activities of scientific practice.

Recognizing the role of context in shaping behavior raises a need for research that examines how students' practices can be altered by enriching learning contexts. Rather than assume that students need to develop knowledge and skills, educators and curriculum designers might consider how a shift in a learning environment might make it possible for students to use existing knowledge and skills in new ways.

Limitations and Open Questions

We have suggested that our design of a laboratory unit impacted students' framings of writing lab reports. In line with other framing research, we inferred students' framings from their behaviors (Scherr and Hammer, 2009). Specifically, we used markers in students' writing to make inferences about students' possible framings of the activity of writing their lab reports. Our inferences could be strengthened with additional observations or interviews with students that could provide more information about how students were framing the lab generally and the lab report specifically.

We also claimed that our design may have functioned to change the relationship between students and instructors. This claim is qualified, because we did not collect data on how implementation of the curriculum and instructional changes varied across lab sections taught by different GTAs. It is possible, for example, that some GTAs were more successful in communicating that they valued students' thinking than others, leaving open questions about the relative role of curriculum and instructor behavior in the framing process.

ACKNOWLEDGMENTS

This work was supported by funding from the Davis Educational Foundation and seed funding as well as an Open Access publishing grant from Tufts University. We wish to thank Robert Hayes and Matt Simon for their design work and discussions of this research. We also thank the editor and two anonymous reviewers for feedback that improved this article.

REFERENCES

American Association for the Advancement of Science. (2011). *Vision and change in undergraduate biology education*. Washington, DC.

Bazerman, C. (1988). *Shaping written knowledge*. Madison, WI: University of Wisconsin Press. https://doi.org/10.1126/science.248.4957.877

Bazerman, C. (2018). What does a model model? And for whom? *Educational Psychologist*, *53*(4), 301–318. https://doi.org/10.1080/00461520.2018.1496022

Berland, L. K., & Hammer, D. (2012). Framing for scientific argumentation. *Journal of Research in Science Teaching*, *49*(1), 68–94. https://doi.org/10.1002/tea.20446

Berland, L. K., & Reiser, B. J. (2011). Classroom communities' adaptations of the practice of scientific argumentation. *Science Education*, *95*(2), 191–216. https://doi.org/10.1002/sce.20420

Biggs, J., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy (structure of the observed learning outcome)*. New York, NY: Academic Press.

Bitzer, L. F. (1968). The rhetorical situation. *Philosophy and Rhetoric*, *1*(1), 1–15.

Blikstein, P., Fuhrmann, T., & Salehi, S. (2016). Using the bifocal modeling framework to resolve "Discrepant Events" between physical experiments and virtual models in biology. *Journal of Science Education and Technology*, *25*(4), 513–526. https://doi.org/10.1007/s10956-016-9623-7

Chen, Y., Benus, M. J., & Hernandez, J. (2019). Managing uncertainty in scientific argumentation. *Science Education*, *103*(5), 1235–1276. https://doi.org/10.1002/sce.21527

Cobb, P., Confrey, J., DiSessa, A. A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher*, *32*(1), 9–13. https://doi.org/10.3102/0013189X032001009

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220.

Engle, R. A., & Conant, F. R. (2002). Guiding principles for fostering productive disciplinary engagement: Explaining an emergent argument in a community of learners classroom. *Cognition and Instruction*, *20*(4), 399–483. https://doi.org/10.2307/3233901

Ford, M. J. (2005). The game, the pieces, and the players: Generative resources from two instructional portrayals of experimentation. *Journal of the Learning Sciences*, *14*(4), 449–487. https://doi.org/10.1207/s15327809jls1404_1

Ford, M. J. (2012). A dialogic account of sense-making in scientific argumentation and reasoning. *Cognition and Instruction*, *30*(3), 207–245. https://doi.org/10.1080/07370008.2012.689383

Ford, M. J. (2015). Educational implications of choosing "practice" to describe science in the Next Generation Science Standards. *Science Education*, *99*(6), 1041–1048. https://doi.org/10.1002/sce.21188

Goffman, E. (1986). *Frame analysis*. Northeastern University Press.

Gouvea, J., & Wagh, A. (2018). Exploring the unknown: Supporting students' navigation of scientific uncertainty with coupled methodologies. In *Proceedings of the 13th International Conference of the Learning Sciences* (pp. 33–40).

Gouvea, J. S., Wagh, A., Hayes, R., & Simon, M. R. (2022). Hybrid labs: How students use computer models to motivate and make meaning from experiments. In Pelaez N. J., Gardner S. M., & Anderson T. (Eds.), *Trends in teaching experimentation in the life sciences. Contributions from biology education research* (pp. 395–413). Switzerland: Springer Nature. https://doi.org/10.1007/978-3-030-98592-9_18

Hammer, D., Elby, A., Scherr, R. E., & Redish, E. F. (2005). Resources, framing, and transfer. In Mestre, J. P. (Ed.), *Transfer of learning from a modern multidisciplinary perspective* (pp. 89–120). Greenwich, CT: Information Age.

Hester, S. D., Nadler, M., Katcher, J., Elfring, L. K., Dykstra, E., Rezende, L., & Bolger, M. S. (2018). Authentic Inquiry through Modeling in Biology (AIM-Bio): An introductory laboratory curriculum that increases undergraduates' scientific agency and skills. *CBE—Life Sciences Education*, *17*(4), ar63. https://doi.org/10.1187/cbe.18-06-0090

Holmes, N. G., & Bonn, D. A. (2014). Doing science or doing a lab? Engaging students with scientific reasoning during physics lab experiments. In *2013 Physics Education Research Conference Proceedings* (pp. 185–188). https://doi.org/10.1119/perc.2013.pr.034

Hu, D., Zwickl, B. M., Wilcox, B. R., & Lewandowski, H. J. (2017). Qualitative investigation of students' views about experimental physics. *Physical Review Physics Education Research*, *13*(2), 1–12. https://doi.org/10.1103/PhysRevPhysEducRes.13.020134

Jiménez-Aleixandre, M. P., Rodríguez, A. B., & Duschl, R. A. (2000). "Doing the lesson" or "doing science": Argument in high school genetics. *Science Education*, *84*(6), 757–792. https://doi.org/10.1002/1098-237X(200011)84:6%3C757::AID-SCE5%3E3.0.CO;2-F

Kelly, G. J., & Bazerman, C. (2003). How students argue scientific claims: A rhetorical-semantic analysis. *Applied Linguistics*, *24*(1), 28–55. https://doi.org/10.1093/applin/24.1.28

Kelly, G. J., Bazerman, C., Skukauskaite, A., & Prothero, W. (2010). Rhetorical features of student science writing in introductory university oceanography. In Bazerman C., Krut R., Lunsford K., McLeod S., Null S., Rogers P., & Stansell A. (Eds.), *Traditions of writing research* (pp. 265–282). New York, NY: Routledge.

Keys, C. W. (1999). Revitalizing instruction in scientific genres: Connecting knowledge production with writing to learn in science. *Science Education*, *83*(2), 115–130.

Kuhn, D., & Pease, M. (2008). What needs to develop in the development of inquiry skills? *Cognition and Instruction*, *26*(4), 512–559. https://doi.org/10.1080/07370000802391745

Kuhn, D., & Udell, W. (2003). The Development of argument skills. *Child Development*, *74*(5), 1245–1260. https://doi.org/10.1111/1467-8624.00605

Lawson, A. E. (2002). Sound and faulty arguments generated by preservice biology teachers when testing hypotheses involving unobservable entities. *Journal of Research in Science Teaching*, *39*(3), 237–252. https://doi.org/10.1002/tea.10019

Lee, H.-S., Liu, O. L., Pallant, A., Roohr, K. C., Pryputniewicz, S., & Buck, Z. E. (2014). Assessment of uncertainty-infused scientific argumentation. *Journal of Research in Science Teaching*, *51*(5), 581–605. https://doi.org/10.1002/tea.21147

Lee, H. S., Pallant, A., Pryputniewicz, S., Lord, T., & Liu, O. L. (2017). Articulating uncertainty attribution as part of critical epistemic practice of scientific argumentation. In Smith B. K., Borge M., Emma Mercier E., & Lim K. Y. (Eds.), *Proceedings of the 12th International Conference on Computer Supported Collaborative Learning*. Philadelphia, PA: International Society of the Learning Sciences.

Libarkin, J., & Ording, G. (2012). The utility of writing assignments in undergraduate bioscience. *CBE—Life Sciences Education*, *11*(1), 39–46. https://doi.org/10.1187/cbe.11-07-0058

Lynch, M. (2010). Evolution of the mutation rate. *Trends in Genetics*, *26*(8), 345–352. https://doi.org/10.1016/j.tig.2010.05.003

Manz, E. (2012). Understanding the codevelopment of modeling practice and ecological knowledge. *Science Education*, *96*(6), 1071–1105. https://doi.org/10.1002/sce.21030

Manz, E. (2015a). Representing student argumentation as functionally emergent from scientific activity. *Review of Educational Research*, *85*(4), 553–590. https://doi.org/10.3102/0034654314558490

Manz, E. (2015b). Resistance and the development of scientific practice: Designing the mangle into science instruction. *Cognition and Instruction*, *33*(2), 89–124. https://doi.org/10.1080/07370008.2014.1000490

Manz, E., Lehrer, R., & Schauble, L. (2020). Rethinking the classroom science investigation. *Journal of Research in Science Teaching*, *57*(7), 1148–1174. https://doi.org/10.1002/tea.21625

McNeill, K. L., Lizotte, D. J., Krajcik, J. S., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *Journal of the Learning Sciences*, *15*(2), 153–191. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.112.2734

Metz, K. E. (2004). Children's understanding of scientific inquiry: Their conceptualization of uncertainty in investigations of their own design. *Cognition and Instruction*, *22*(2), 219–290. https://doi.org/10.1207/s1532690xci2202

Moon, A., Stanford, C., Cole, R., & Towns, M. (2017). Analysis of inquiry materials to explain complexity of chemical reasoning in physical chemistry students' argumentation. *Journal of Research in Science Teaching*, *54*(10), 1322–1346. https://doi.org/10.1002/tea.21407

Moskovitz, C., & Kellogg, D. (2011). Inquiry-based writing in the laboratory course. *Science*, *332*(6032), 919–920.

National Research Council. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies Press. www.nap.edu/catalog.php?record_id=13165

O'Neill, D. K. (2001). Knowing when you've brought them in: Scientific genre knowledge and communities of practice. *Journal of the Learning Sciences*, *10*(3), 223–264. https://doi.org/10.1207/S15327809JLS1003

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching*, *41*(10), 994–1020. https://doi.org/10.1002/tea.20035

Osborne, J., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., & Yao, S. Y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching*, *53*(6), 821–846. https://doi.org/10.1002/tea.21316

Petraglia, J. (1995). Spinning like a kite: A closer look at the pseudotransactional function of writing. *JAC*, *15*(1), 19–33.

Petritis, S. J., Kelley, C., & Talanquer, V. (2021). Exploring the impact of the framing of a laboratory experiment on the nature of student argumentation. *Chemistry Education Research and Practice*, *22*(1), 122–135. https://doi.org/10.1039/d0rp00268b

Quitadamo, I. J., & Kurtz, M. J. (2007). Learning to improve: Using writing to increase critical thinking performance in general education biology. *CBE—Life Sciences Education*, *6*, 140–154. https://doi.org/10.1187/cbe.06

Russ, R. S., & Berland, L. K. (2019). Invented science: A framework for discussing a persistent problem of practice. *Journal of the Learning Sciences*, *28*(3), 279–301. https://doi.org/10.1080/10508406.2018.1517354

Ryu, S., & Sandoval, W. a. (2012). Improvements to elementary children's epistemic understanding from sustained argumentation. *Science Education*, *96*(3), 488–526. https://doi.org/10.1002/sce.21006

Sandoval, W. A. (2014). Conjecture mapping: An approach to systematic educational design research. *Journal of the Learning Sciences*, *23*(1), 18–36. https://doi.org/10.1080/10508406.2013.778204

Sandoval, W. A., & Millwood, K. A. (2005). The quality of students' use of evidence in written scientific explanations. *Cognition and Instruction*, *23*(1), 23–55. https://doi.org/10.1207/s1532690xci2301_2

Schen, M. (2017). A comparison of biology majors' written arguments across the curriculum. *Journal of Biological Education*, *47*(4), 224–231. https://doi.org/10.1080/00219266.2013.788542

Schank, R., & Abelson, R. (1977). *Script, plans, goals and understanding*. Hillsdale, NJ: Lawrence Erlbaum.

Scherr, R. E., & Hammer, D. (2009). Student behavior and epistemological framing: Examples from collaborative active-learning activities in physics. *Cognition and Instruction*, *27*(2), 147–174. https://doi.org/10.1080/07370000902797379

Schwarz, B. B., Neuman, Y., Gil, J., & Ilya, M. (2003). Construction of collective and individual knowledge in argumentative activity. *Journal of the Learning Sciences*, *12*(2), 219–256. https://doi.org/10.1207/S15327809JLS1202

Smith, E. M., Stein, M. M., & Holmes, N. G. (2020). How expectations of confirmation influence students' experimentation decisions in introductory labs. *Physical Review Physics Education Research*, *16*(1), 10113. https://doi.org/10.1103/PhysRevPhysEducRes.16.010113

Stein, M. M., Smith, E. M., & Holmes, N. G. (2018). Confirming what we know: Understanding questionable research practices in intro physics labs.

Physics Education Research Conference Proceedings, *2018*, 1–4. https://doi.org/10.1119/perc.2018.pr.stein

Tang, X., Coffey, J. E., & Levin, D. M. (2015). Reconsidering the use of scoring rubrics in biology instruction. *American Biology Teacher*, *77*(9), 669–675. https://doi.org/10.1525/abt.2015.77.9.4.THE

Tannen, D. (1979). What's in a frame? Surface evidence for underlying expectations. In Freedle, R. (Ed.), *New directions in discourse processing* (pp. 137–181).. Norwood, NJ: Ablex.

Tannen, D., & Wallat, C. (1987). Interactive frames and knowledge schemas in interaction: Examples from a medical examination/interview. *Social Psychology Quarterly*, *50*(2), 205. https://doi.org/10.2307/2786752

Walker, J. P., & Sampson, V. D. (2013). Argument-driven inquiry: Using the laboratory to improve undergraduates' science writing skills through meaningful science writing, peer-review, and revision. *Journal of Chemical Education*, *90*, 1269–1274. https://doi.org/10.1021/ed300656p

Xu, H., & Talanquer, V. (2013). Effect of the level of inquiry of lab experiments on general chemistry students' written reflections. *Journal of Chemical Education*, *90*(1), 21–28. https://doi.org/10.1021/ed3002368

Zagallo, P., Meddleton, S., & Bolger, M. S. (2016). Teaching Real Data Interpretation with Models (TRIM): Analysis of student dialogue in a large-enrollment cell and developmental biology course. *CBE—Life Sciences Education*, *15*(2), ar17. https://doi.org/10.1187/cbe.15-11-0239