2015 Highlights

The journal is supported in part by a grant from the Howard Hughes Medical Institute
2015 HIGHLIGHTS

Table of Contents

EDITORIAL

Biology Education Research 2.0
Erin L. Dolan ................................................................. 1–2

Biology education research (BER) 2.0 has arrived, and is moving the BER community beyond showing that active learning works to understanding the individual and contextual factors that explain and influence biology teaching and learning.

LETTER TO THE EDITOR

Overcoming the Barrier to Implementing Authentic Research Experiences through Faculty Mentorship
Christine M. Goedhart and Jacqueline S. McLaughlin ......................................................... 3–4

FEATURES

Approaches to Biology Teaching and Learning
A Portal into Biology Education: An Annotated List of Commonly Encountered Terms
Sarah Miller and Kimberly D. Tanner ................................................................. 5–18

Exploring a new discipline can be daunting in any field, and biology education is no exception. The authors provide a resource for those who are new to explorations of the biology education and biology education research worlds, including key terminology, brief definitions, and links to literature for further explorations.

WWW.Life Sciences Education

Online Resources for Understanding Outbreaks and Infectious Diseases
Nicola C. Barber and Louisa A. Stark ................................................................. 19–23

Disease outbreaks can be a powerful topic for teaching about science and health. This Feature reviews resources for bringing up-to-date information on this hot topic into the classroom.

Current Insights

Recent Research in Science Teaching and Learning
Deborah Allen ................................................................. 25–27

This feature is designed to point CBE—Life Sciences Education readers to current articles of interest in life sciences education as well as more general and noteworthy publications in education research.
Book Review

Integrating Concepts in Biology: A Model for More Effective Ways to Introduce Students to Biology
K. N. Prestwich and A. M. Sheehy ................................................................. 29–32

This is a review of Integrating Concepts in Biology, an innovative electronic textbook intended for either or both semesters of a typical introductory biology sequence. The e-book is an excellent example of effective pedagogy informed by recent educational scholarship.

Meeting Report

NEST 2014: Views from the Trainees—Talking About What Matters in Efforts to Diversify the STEM Workforce
Andrew G. Campbell, Rachel Skvirsky, Henry Wortis, Sheila Thomas, Ichiro Kawachi, and Christine Hohmann ................................................................. 33–38

This paper summarizes the outcomes of a retreat designed to cultivate interactions between trainees at various training levels and provide them opportunities to share their training perspectives and expectations. Retreat outcomes are used to support the development of better science, technology, engineering, and mathematics training practices by informing the trainers’ perspective.

Essays

Research-Based Implementation of Peer Instruction: A Literature Review
Trisha Vickrey, Kaitlyn Rosploch, Reihaneh Rahmanian, Matthew Pilarz, and Marilyne Stains ............ 39–49

Peer instruction is an evidence-based pedagogy that has been extensively studied in various science, technology, engineering, and mathematics disciplines. In this essay, the authors review and summarize the research literature on the effectiveness and intricacies of implementation of peer instruction. A research-based how-to guide and suggestions for future research investigations are provided.

Modeling Course-Based Undergraduate Research Experiences: An Agenda for Future Research and Evaluation
Lisa A. Corwin, Mark J. Graham, and Erin L. Dolan................................................................. 51–63

The authors review relevant literature to determine established and predicted outcomes of course-based undergraduate research experiences (CUREs) and then use this information and social learning theory to model how students may realize desired short-, medium-, and long-term outcomes. This work has implications for future research and evaluation of CUREs.

Articles

Considering Student Voices: Examining the Experiences of Underrepresented Students in Intervention Programs
Gina Sanchez Gibau ................................................................. 65–76

This article presents a qualitative analysis of the effectiveness of intervention programs designed to address underrepresentation in the biomedical sciences. The article highlights the perspectives and experiences of underrepresented students participating in such programs.
Helping Struggling Students in Introductory Biology: A Peer-Tutoring Approach That Improves Performance, Perception, and Retention

This study examines an optional peer-tutoring program offered to students who are struggling in a large-enrollment, introductory biology course. Students who regularly attended had increased exam performance, more expert-like perceptions of biology, and increased persistence relative to their struggling peers who were not attending.

Use of Feedback-Oriented Online Exercises to Help Physiology Students Construct Well-Organized Answers to Short-Answer Questions
Jacqueline Carnegie .................................................................89–100

Online feedback-oriented exercises were developed to help undergraduate physiology students practice formulating answers to short-answer questions. Student feedback regarding the learning value of these assignments and the ability of these assignments to improve student outcomes when answering short-answer questions on summative examinations were assessed.

Differences in Metacognitive Regulation in Introductory Biology Students: When Prompts Are Not Enough
Julie Dangremond Stanton, Xyanthe N. Neider, Isaura J. Gallegos, and Nicole C. Clark ...............101–112

Metacognition correlates with learning outcomes and student performance. In this study, the authors examined the metacognitive-regulation skills used by introductory biology students. They found that prompting students to use these skills is effective for some students, but other students need additional help with learning strategies to respond optimally.

The Synthesis Map Is a Multidimensional Educational Tool That Provides Insight into Students’ Mental Models and Promotes Students’ Synthetic Knowledge Generation
Ryan A. Ortega and Cynthia J. Brame ........................................113–123

Synthesis mapping uses a novel, multidimensional presentation tool to allow students to create a detailed map of their hierarchal knowledge structures. It can be used as formative assessment to reveal students’ strengths, misconceptions, and organizational schema and as a summative assessment to test students’ understanding of course material.

Relations between Intuitive Biological Thinking and Biological Misconceptions in Biology Majors and Nonmajors
John D. Coley and Kimberly Tanner .........................................125–143

The authors present evidence that seemingly unrelated biological misconceptions may share common conceptual origins arising from underlying systems of intuitive biological reasoning, or “cognitive construals.” The findings presented raise the intriguing possibility that university-level biology education may reify construal-based thinking and related misconceptions.

A High-Enrollment Course-Based Undergraduate Research Experience Improves Student Conceptions of Scientific Thinking and Ability to Interpret Data
Sara E. Brownell, Daria S. Hekmat-Scafe, Veena Singla, Patricia Chandler Seawell, Jamie F. Conklin Imam, Sarah L. Eddy, Tim Stearns, and Martha S. Cyert ..........................145–158

The authors developed and assessed an innovative course-based undergraduate research experience that emphasized collaboration among students and focused on data analysis.
PORTAAL: A Classroom Observation Tool Assessing Evidence-Based Teaching Practices for Active Learning in Large Science, Technology, Engineering, and Mathematics Classes
Sarah L. Eddy, Mercedes Converse, and Mary Pat Wenderoth .................................159–174

PORTAAL, a new evidence-based classroom observation tool, identifies 21 elements of classroom best practices for active learning that have been correlated with positive student outcomes in the education literature. After only 5 h of training, instructors can reliably use this tool to determine their alignment with these teaching practices.

Information for Authors .................................................................175–179
Editorial

Biology Education Research 2.0

Erin L. Dolan

Texas Institute for Discovery Education in Science, College of Natural Sciences, University of Texas at Austin, Austin, TX 78712

Earlier this year, I was enjoying dinner with an influential colleague at an education conference. As would be expected given the venue, this colleague was interested in good teaching. I assumed she would be familiar with the abundance of research on the effectiveness of active learning (e.g., Hake, 1998; Prince, 2004; Ruiz-Primo et al., 2011; Freeman et al., 2014). After listening to one of our tablemates describe an active-learning strategy he had used, my companion turned to me and asked, “But how do we know that it works?” Her question caught me by surprise. Hadn’t this question been answered time and time again?

This moment of disconnect between research on effective instruction and the many instructors and other decision makers who are not yet familiar with this body of knowledge got me thinking about the power of the CBE—Life Sciences Education (LSE) community. Since the inception of the journal, LSE editors, authors, and readers have been called on to serve as translators between what is known about teaching and learning and how teaching is practiced. LSE has put this translational role into practice in a number of ways, such as the Approaches to Biology Teaching and Learning and Current Insights features and the Research Methods essays.

These, along with many other articles and essays published in LSE, have described or cited the guiding principles for teaching that promotes active learning: engaging students, aiming for an outcome or objective, providing structure and opportunities for practice, giving feedback, encouraging interaction and reflection, expecting higher-level thinking, informing instructional decisions with evidence of student learning and development, and incorporating well-motivated and well-timed explanations from reading or mini-lectures (Bransford et al., 1999; Singer et al., 2012; Dolan and Collins, 2015; Kober, 2015). We know this type of teaching works when deployed well (Freeman et al., 2014), and that it works especially well for students who have been traditionally underserved (e.g., Eddy and Hogan, 2014).

Imagine for a moment that teaching using active learning is a construction project, and the goal is to construct student learning. The construction tools (e.g., screwdriver, hammer) are the instructional materials (e.g., assignments, clicker questions, exams), and how the tools are used is the instructional strategy. At this point in understanding teaching and learning, we know how to design the screwdriver and the hammer. We also know how a screwdriver and a hammer should be used, and that some aspects of construction will require a screwdriver, while others will require a hammer.

A person who is new to construction may not know that the hammer, rather than the handle of the screwdriver, is a better tool to drive in a nail. He or she may not know that a particular screw requires the use of a Phillips-head instead of a flathead screwdriver. This does not mean we need to redemonstrate that a screwdriver or hammer works. Rather, we need to figure out ways to help all involved in construction to learn how useful the tools are, how to select the right tools for the job, how to use the tools, and what latitude there is for using a range of tools.

This is the direction in which we need to head with the study of biology education. We need to know what is happening during active learning that makes it work—at the levels of the student, instructor, discipline, department, and institution. We need to understand what working means, for whom, and in what contexts (Tanner, 2011). This will require a different kind of research—what some are calling the next generation of biology education research (BER), or BER 2.0.

Excitingly, the LSE community is already making progress in this direction. Several recent articles in the journal have aimed at demonstrating what makes “flipped instruction” work (Gross et al., 2015; Jensen et al., 2015) and what “working” means (van Vliet et al., 2015). To continue to make progress in this direction, we need to look to other fields for theory and methods, including cognitive science, psychology, sociology, and anthropology, while keeping in mind our important role in translating the work in these fields, so it is comprehensible to a much broader audience. We need to think creatively about how to bring life sciences research methods—such as those used to study physiological systems, to model ecological processes across scales, and to analyze metabolic networks—to bear on the study of teaching and learning. We need to examine research from such diverse environments as K–12 education and corporate settings and to envision how it might help us understand biology education at other levels and in other settings.
We have embraced concept inventories to measure student learning, which has been an important driver of deeper consideration of how we assess our students’ learning. Now we need to explore other ways of thinking about student cognition (Pellegrino et al., 2001), such as threshold concepts (Meyer, 2008; Meyer and Land, 2006; Loertscher et al., 2014); learning progressions (Alonzo and Gotwals, 2012); and schema, phenomenological primitives, and cognitive construals (diSessa, 1988, 1993; Hammer, 1996; Coley and Tanner, 2015). We need to examine other ways biology learners develop, for example, in their identities as scientists; their sense of belonging to science; or their abilities to reflect, self-regulate, and embrace a growth mind-set (Duckworth and Yeager, 2015). We need to balance our need to use common instruments to compare results across studies with our need to develop new and better ways to measure important outcomes that will help us improve the experience of learning biology (e.g., Pellegrino et al., 2001; Yeager et al., 2013).

We need to study instructional change beyond single classrooms or institutions. For example, how do faculty develop knowledge and skills important for teaching research courses (Auchincloss et al., 2014), supporting all students in learning, or guiding students in learning particular domains of life science? What lessons learned from professional development in other disciplines and K–12 settings apply to understanding experiences of biology faculty? In what ways do our institutions differ in their teaching climates, cultures, operations, and incentive systems, and how do these differences support or constrain faculty members in improving their teaching? Again, we can inform our research in these areas by exploring other fields, such as industrial and organizational psychology, improvement science, and health systems research (Campbell et al., 2000; World Health Organization, 2015).

BER 2.0—moving beyond answering the question of whether it works—will be best positioned to thrive if we continue to embrace our role in translating what is known about teaching and learning so that it can both inform our work and serve a broader audience of biology educators. This has been priority and a defining feature of LSE since its inception and will be the focus of a new phase of development of the journal in 2016. Stay tuned!

REFERENCES
Letter to the Editor

Overcoming the Barrier to Implementing Authentic Research Experiences through Faculty Mentorship

Christine M. Goedhart* and Jacqueline S. McLaughlin†

*Department of Biology, Citrus College, Glendora, CA 91741; †Department of Biology, Pennsylvania State University, Lehigh Valley, Center Valley, PA 18034

To the Editor:

The American Association for the Advancement of Science’s Vision and Change report called for authentic research experiences for all undergraduate biology students specifically for the purpose of training students to gain the skills required to face, and solve, our current and future problems (American Association for the Advancement of Science, 2011). However, such experiences are rare, particularly in introductory biology courses. Spell et al. (2014) recently put forth the results of a national survey in which instructors reported that undergraduate students in introductory biology courses spend on average only one-third of their time on authentic research activities, with 23% of courses having no research and 56% spending less than a quarter of total class time engaging in research. Instructors in their study also reported that the amount of time required to develop research experiences was the most common barrier to implementing them, regardless of institution type (Spell et al., 2014).

We suggest that forming mentoring partnerships between instructors who have already developed research experiences and instructors who wish to do so is an effective way to overcome the time barrier. We formed such a partnership and found it saved time otherwise spent reinventing the wheel and also offered benefits such as encouragement, accountability, example (shared) instructional materials and assessment tools, pedagogical strategies to enhance student engagement in primary literature and writing skills, and camaraderie.

We (a four-year university mentor [J.S.M.] and a two-year college mentee [C.M.G.]) launched our partnership in 2014, meeting monthly by means of various technologies (phone, video conference, and Skype). During the initial meetings, J.S.M. mentored C.M.G. in the framework of the guided inquiry–based research model created by J.S.M. We then identified a semester and course in which C.M.G. would implement the research model, and we developed an overall timeline and schedule to achieve this goal. We worked together to adapt the model to work within a two-year college nonmajors biology course. Each meeting involved the sharing of progress reports and the creation of future action plans. J.S.M. also shared course materials, discussed personal experiences, and taught C.M.G. how to use particular assessment tools and teaching strategies. By the end of the partnership, C.M.G. had successfully implemented the guided inquiry–based research experience in a nonmajors biology course.

This partnership was initiated following J.S.M.’s presentation at the 2013 National Association of Biology Teachers annual conference. As an audience member during the presentation, C.M.G. identified J.S.M.’s research experience model as one that she wanted to implement in her own course and reached out to J.S.M. for guidance. This example highlights the importance of live meeting venues where instructors can share and present research experience models they have developed and implemented with success. These types of presentations often occur at professional society meetings, making these meeting venues the ideal setting to promote the formation of professional mentoring partnerships. Special symposia could be developed for the specific purpose of bringing together potential mentors and mentees. Instructors who are willing to serve as mentors would have the opportunity to briefly present their research experience model to an audience of instructors who are seeking mentorship.

As we found in our partnership, research experience models do not always transfer smoothly from one institution to another. For example, C.M.G. could not directly implement J.S.M.’s research topic, which was developed within a four-year university because the necessary equipment and infrastructure was not available at C.M.G.’s two-year college. Even in its adapted form, the research experience model required the purchase of additional equipment and supplies. Therefore, to effectively facilitate the transfer of successful research experience models, grant programs could be developed to provide funding for equipment and supplies needed.
to transfer successful research experience models to a variety of courses and institutions.

We assert that a mentoring partnership is an effective way to overcome the barrier of time on the side of the mentee. However, this is not the case for the mentor. Mentoring requires time, patience, and a willingness to share materials. To make the mentoring role more meaningful and worthwhile, it should be validated and included within promotion and tenure policies as professional development and as service to the educational community. Exemplars, showcased in feature articles in biology peer-reviewed education journals, would validate the importance of this service and allow mentors and mentees the ability to professionally reference their work. In closing, professional mentoring partnerships may be an important catalyst for increasing authentic research experiences in all biology courses nationwide while working to infuse the directives of Vision and Change at the institutional level and properly acknowledging a core academic and commonly understated ingredient, mentorship.

ACKNOWLEDGMENTS

We thank Melissa Coyle for her role in helping to develop the guided inquiry–based research laboratory model at Penn State Lehigh Valley, for providing expert lab technical support, and for helping to coordinate the mentoring partnership logistics.

REFERENCES


Feature

Approaches to Biology Teaching and Learning

A Portal into Biology Education: An Annotated List of Commonly Encountered Terms

Sarah Miller* and Kimberly D. Tanner†

*Faculty Engagement Services, Division of Information Technology, Academic Technology, University of Wisconsin–Madison, Madison, WI 53706; †Department of Biology, San Francisco State University, San Francisco, CA 94132

In an introductory biology course, undergraduate students are expected to become familiar with, and be able to use, hundreds of new terms to navigate the complex ideas in biology. And this is just in the introductory course! Juxtapose this student situation with the common frustration expressed by biologists that there is “just too much jargon in science education.” Unfortunately, a common frustration for a disciplinary novice is learning to navigate the language. For biologists venturing into the social sciences, this can be particularly tricky, as the language may at first seem easily understood, because the conjoined words are familiar—for example, “cooperative learning,” “stereotype threat,” or “problem-based learning.” However, the meaning of these words in combination is often specialized and represents rich traditions and research literatures that go far beyond summing the dictionary definitions of the component terms.

So, how does one begin to explore the ideas and language of biology education? We have chosen 50 key terms that scientists will likely encounter in any exploration of biology education. To provide a framework for how these terms might connect together for instructors, we have used the organizing framework of scientific teaching, in which there is no prescribed or correct way to teach; rather, instructors are expected to apply scientific principles to their classroom teaching efforts. Scientific teaching is an intentional approach to teaching by instructors that focuses on the goal of student learning and involves iterative questioning, evidence collection, and innovation. Inspired by the original book on the subject, Scientific Teaching (Handelsman et al., 2006), we have chosen to introduce the reader to key terminology in biology education by organizing these terms with respect to the three main tenets of scientific teaching—active learning; assessment; and the related ideas of equity, diversity, and inclusivity—along with a fourth section about tools for moving the ideas of scientific teaching into practice:

Active Learning: Engaging Students as Participants in Learning Assessment: Finding Out How Students Are Thinking and Learning Equity, Diversity, and Inclusivity: Creating Fair and Accessible Learning Environments Moving to Practice: Instructional Design, Learning, and Technologies

For each of these four sections, there is a brief overview of the topic, followed by a set of commonly encountered terms related to that topic. For each key term, we provide an introductory, descriptive paragraph, which is followed by two references that could be starting points for additional explorations. Whenever possible, these references include accessible review articles written primarily for a scientific audience. No doubt, dozens of additional terms could be added to each section; however, this collection is intended to be a starting point for readers. Additionally, many of the terms we have associated with one section could easily also be placed into other sections. For example, “think–pair–share” is a teaching strategy that is all at once an active-learning approach, a potential mode to assess students, and an equitable teaching strategy that can provide access and opportunity for all students to participate.

Importantly, the entries for these 50 key biology education terms may be read and explored in any order. Readers are encouraged to use Table 1 to self-assess which of these terms are familiar and most interesting. Self-assessment responses to Table 1 can guide which of the sections below you may be most interested in reading first. So, onward, and enjoy exploring the ideas and language of biology education.
ACTIVE LEARNING: ENGAGING STUDENTS AS PARTICIPANTS IN LEARNING

The first cornerstone of scientific teaching is active learning. As a biology education community, we have traditionally focused much of our conversation about teaching and learning on issues of “what” exactly students should be learning; however, attention is increasingly being paid to the “how” of teaching. Multiple lines of research efforts in a variety of disciplines have provided evidence that traditional lecture approaches to teaching are much less effective than teaching approaches that actively engage students in the learning process (Bransford et al., 1999; Handelsman et al., 2006; Freeman et al., 2014). Active-learning approaches to teaching encompass a range of strategies—from simple to complex, from activities that last just a few minutes to longer projects,

Table 1. Biology education terminology self-assessment

<table>
<thead>
<tr>
<th>U=Unfamiliar</th>
<th>F=Familiar</th>
<th>Biology Education Terminology Self-Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td><strong>Active Learning</strong></td>
</tr>
<tr>
<td><em>1. Think-Pair-Share</em></td>
<td><em>12. Formative Assessment</em></td>
<td></td>
</tr>
<tr>
<td><em>2. Clickers</em></td>
<td><em>13. Summative Assessment</em></td>
<td></td>
</tr>
<tr>
<td><em>4. Group Work</em></td>
<td><em>15. Postassessment</em></td>
<td></td>
</tr>
<tr>
<td><em>5. Cooperative Learning</em></td>
<td><em>16. Learning Gains</em></td>
<td></td>
</tr>
<tr>
<td><em>6. Peer Instruction</em></td>
<td><em>17. Closed-ended Questions</em></td>
<td></td>
</tr>
<tr>
<td><em>10. Student Resistance</em></td>
<td><em>21. Grading</em></td>
<td></td>
</tr>
<tr>
<td><em>11. Student Evaluation</em></td>
<td><em>22. Criterion-referenced Grading</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Inclusivity, Equity, and Diversity</strong></td>
</tr>
<tr>
<td><em>25. Equity</em></td>
<td><em>40. Backward Design</em></td>
<td></td>
</tr>
<tr>
<td><em>27. Inclusivity</em></td>
<td><em>42. Metacognition</em></td>
<td></td>
</tr>
<tr>
<td><em>28. Student-Deficit Model</em></td>
<td><em>43. Syllabus</em></td>
<td></td>
</tr>
<tr>
<td><em>29. Achievement Gap</em></td>
<td><em>44. Lesson Plans</em></td>
<td></td>
</tr>
<tr>
<td><em>30. Instructional Selection</em></td>
<td><em>45. Learning Technologies</em></td>
<td></td>
</tr>
<tr>
<td><em>31. Stereotype Threat</em></td>
<td><em>46. Blended Learning</em></td>
<td></td>
</tr>
<tr>
<td><em>32. Retention</em></td>
<td><em>47. Accessibility</em></td>
<td></td>
</tr>
<tr>
<td><em>33. Persistence</em></td>
<td><em>48. Universal Design</em></td>
<td></td>
</tr>
<tr>
<td><em>34. Underrepresented Minority</em></td>
<td><em>49. Scientific Teaching</em></td>
<td></td>
</tr>
<tr>
<td><em>35. Students of Color</em></td>
<td><em>50. Nature of Science</em></td>
<td></td>
</tr>
<tr>
<td><em>36. First-Generation Students</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>37. Self-Efficacy</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>38. Unconscious or Implicit Bias</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>39. Equitable Teaching Strategies</em></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Other biology education terms I have encountered and want to explore...
from inside class time to outside class time. Common to all these active-learning strategies is the acknowledgment that learning is a phenomenon of the human brain, and the individuals doing the learning must be actively involved in constructing meaning, examining their prior ideas, and resolving conceptual confusions, just as scientists do in their own efforts to learn how the natural world works. Below, we introduce 11 key terms that biologists would likely encounter in their explorations to learn more about active learning.

1. **Think–Pair–Share**

The term “think–pair–share” refers to a teaching method that expects students think individually about a solution to a problem for a moment, then pair with a neighbor to share their ideas, and sometimes eventually report out to the large group (Lyman, 1981; reviewed in Tanner and Allen, 2002). In three easy steps, every student in a class of any size can be engaged in active learning through a think–pair–share. After posing a question, the instructor gives the class a few minutes to think and jot down their thoughts. This think time is key, since different students may have different cognitive processing times—our brains all work differently—and giving students more time to just think has been shown to increase the quality of comments later shared and the number of students willing to share (Rowe, 1987; Tobin, 1987). Then comes the “pair” time, a few minutes for each student to say his or her ideas out loud to another student in the class. For the vast majority of students who do not have the confidence to ask or answer questions in front of the whole class, this pair time may be the first time they have uttered a word in an undergraduate science classroom. Pair time allows students to articulate their ideas in the presence of another person; compare their ideas with those of a peer; and identify points of agreement, disagreement, and confusion. Finally comes the “share” part, in which several students are asked to share with the whole class ideas that emerged in their pair discussions. This phase should be very familiar to most instructors, since it is comparable to posing a question to an entire class. Setting up a think–pair–share activity can be as simple as posing a question or problem for students to think about and discuss, such as “Predict the outcome of this experiment,” “Propose at least two hypotheses to explain these observations,” or “Answer the multiple-choice question posted on this slide.”

For further exploration...


2. **Clickers**

Clickers—also referred to as personal response systems—are devices that can be used in classrooms of any size to ask multiple-choice questions with the goal of engaging students with the course material as part of active-learning exercises. While a variety of clicker systems are available, the iClicker system has become widespread, likely because these clickers do not require integration with a particular software presentation system. While clickers can be used to check attendance, more effective uses of clickers aim to engage students in answering questions that check student understanding, challenge common misconceptions, and provide immediate conceptual feedback for both students and instructors alike. While getting students to talk does not require clickers, clicker questions can be the basis of multiple think–pair–share activities during a single class period. A key added value of using clickers is that this technological tool can give the instructor an instant summary of the distribution of student responses to a multiple-choice question. This information can immediately guide an instructor in deciding how to proceed, depending on the proportion of students who select the most scientifically accurate response. Clicker questions can be especially useful when asked a week or so before a section of the course is taught. Using this evidence from students, instructors can identify common misconceptions held by those students, plan class activities to address these misconceptions, spend less time on those ideas that students already seem to know, and, finally, share this clicker evidence with students to explain why course time is being spent on some particular topics more than others.

For further exploration...


3. **Minute Paper**

A minute paper is a brief active-learning strategy that provides a mechanism for students to stop, think, and write during or at the end of a class period (Mazur, 1996). The goal is to provide a momentary break during which students can capture their thoughts or questions. While referred to as a “minute” paper, these brief writings can generally take one to several minutes, depending on the complexity of the question being asked. Often, questions that are most effective at challenging students’ ideas and promoting rich discussions are not multiple choice, in which case clickers become less useful. Minute papers are often driven by these non–multiple choice questions (see 18. *Open-Ended Questions*). Example minute-paper prompts might include: “What’s the most useful concept or idea you learned in class today?,” or “What was the muddiest point in today’s class session that was most confusing for you?,” or “Push yourself to write down at least two questions you have about the scientific evidence we explored in class today.” Some instructors require students to purchase a 100-pack of index cards as part of their course materials to facilitate frequent use of minute papers. Students are told that during each of their class meetings over the semester, they will be asked to write down their ideas on these index cards. Most of the time, instructors will collect these cards, but sometimes they will not. A minute paper
can serve as one way to accomplish the “think” phase of a think–pair–share, since it actively engages students in doing something to drive their thinking.

For further exploration...


4. Group Work and 5. Cooperative Learning

Science is, by nature, a collaborative endeavor, and all scientific careers to which undergraduate students aspire will require extensive skills in working collaboratively. Group work, also referred to as cooperative learning, is a term that refers to activities that require students to engage in active learning with others, during which they work together toward a common outcome and practice improving their collaborative skills (Johnson et al., 1991, 1993, 1998). To be successful, group work requires several critical elements. First, the task must be clear, with students understanding both the final goal of the activity, as well as key milestones along the way. Second, the assignment must be sufficiently complex that it necessitates collaboration. Students are clever; they know that being told to work together on a simple task is not a good use of time and will then perceive the task as busy work. As such, group sizes should reflect the complexity of the assigned task. Third, students need to know their role and the instructor’s expectations of them as individuals. Roles can be assigned by the instructor, or in some special cases, students can be charged with self-organizing. Roles can be divided multiple ways, depending on the task. For example, roles can reflect group functions, such as facilitator, timekeeper/recorder, reporter, or equity monitor (the person who makes sure all group members’ ideas are heard). Roles can also reflect authentic perspectives on a problem or issue, such as scientist, policy analyst, financial officer, business owner, or citizen/parent.

Finally, group work requires the establishment of trust, grounded in a set of group norms, which are guidelines for working together that may be set by the instructor or students in the course. Examples of group norms are that everyone will contribute during discussions, ideas will be respectfully shared, and work will be fairly divided among group members. Group work and cooperative learning is often assumed to include more than two students, but usually no more than six. Importantly, the larger the size of the student group, the more care the instructor must take to clearly define tasks and roles to ensure no students are left out of the group process.

For further exploration...


6. Peer Instruction and 7. Peer-Led Team Learning

The practice of students teaching students is not new and has taken many forms over the years, as teaching an idea to someone else is often an effective way to learn. Peer instruction is defined by the act of students teaching or reviewing one another. In some peer-instruction scenarios, one student becomes an “expert” on some topic and is then tasked to teach what he or she knows to other students who are novices with respect to that material. This approach is often used as an active-learning strategy for exploring large amounts of complex material in single class periods. For example, four research articles can be explored in a single class session in which one-quarter of the class has become expert on each paper through a prior homework assignment. During class time, student experts for each of the four papers then take turns sharing their insights and entertaining questions from peers. Peer instruction also refers to opportunities in classrooms wherein students teach one another without there necessarily being different levels of expertise between students (Mazur, 1996; Gossler et al., 2001; Smith et al., 2009). Research has shown that two students, neither of whom was able to correctly answer a clicker question, could improve their understanding and correctly answer similar clicker questions after engaging in a pair discussion, which is another form of peer instruction (Smith et al., 2009). Peer-led team learning is a particular type of peer instruction, in which students who have previously excelled in a course are invited back to serve as supplementary course discussion leaders or in-class coaches for students currently taking the course. Having most recently learned the concepts at hand, these peers—sometimes called learning assistants—may be more socially accessible to students and better able to remember the kinds of confusions students encounter with the material.

For further exploration...


8. Case-Based Learning and 9. Problem-Based Learning

Case-based and problem-based learning are teaching approaches that link course concepts to real-world scenarios and problems with which students actively engage through exploring and questioning and applying biological content knowledge. While the goal is similar for each, the approaches vary slightly. Problem-based learning moves groups of students through a prescribed strategy of identifying what they already know and what they need to know, then determining how to access any additional information they need to solve a complex problem related to biology, such as determining mechanisms of gene inheritance or designing an experiment to distinguish between alternate hypotheses. Case-based learning evolved from the medical and business fields; the activities task students with figuring out the underlying causes of a medical condition or a business success. Similarly, in science instruction, case-based learning begins with a situation or scenario that poses one or more issues the students need to address. Depending on the goals of the
case, students may be given more or less structure in how to resolve it. Cases can range from fairly simple (one issue, one solution) to complex, real-world scenarios that scientific researchers themselves may not yet have been able to explain. Importantly, both case-based and problem-based learning usually involve students working in structured groups in which they collaboratively identify questions and confusions and seek out additional information to expand their understanding of concepts related to the problem or case.

For further exploration...


10. Student Resistance and 11. Student Evaluations
As a new instructor, student resistance—and the potential for poor student evaluations at the end of the course—can seem sufficient reason to avoid attempting active learning and becoming more innovative in the classroom. Student resistance is defined as unwillingness among one or more students to comply with instructor requests. This resistance can take the form of active resistance, including complaints to department heads or vocal refusal to participate, as well as more passive forms of resistance, such as just not following instructor directions and doing something else. Student resistance often seems to be rooted in students’ prior experiences in classrooms and their expectations of what should be happening in an undergraduate science course. If students have experienced lecture throughout their school years, then their expectations may be unmet by this shift from their normal experiences when they enter a course with extensive active learning.

Student evaluations, defined as information gathered about the course or the teaching at some point during the semester, are a concern for most faculty members. For example, students may have previously experienced poorly structured active learning or group work, resulting in generally unfavorable opinions of these teaching approaches. This is where an instructor’s transparency about the reasons for choosing particular teaching strategies may be critical. In addition to sharing their rationale for teaching choices, instructors can also collect student evaluations throughout the course, not just at the end. This not only provides students a voice in the teaching and learning process but also gives instructors insight into how students are experiencing the course midway, offering the opportunity for the instructor to make adjustments. Using the minute-paper method described earlier, instructors can ask each student anonymously to share responses to: “So far, what aspects of the course are most supporting your learning?” and “So far, what aspects of the course are least supporting your learning?” Even the act of inviting student insights may go far in quelling student resistance. Finally, it may be key for instructors to be quantitative in gauging the extent of student resistance and to be systematic in hearing from all students. While a few students may express resistance, the vast majority of students could deeply appreciate innovative teaching approaches, and the instructor would be unaware of this without inviting midcourse student evaluation.

For further exploration...


ASSESSMENT: FINDING OUT HOW STUDENTS ARE THINKING AND LEARNING
The second tenet of scientific teaching is assessment, which is fundamentally the act of using questions to reveal student thinking, challenge student thinking, and gauge the extent to which we are changing students’ minds about how the biological world works. Importantly, assessment is the act of asking a question and receiving responses, not just from one or a few students, but from all students in a classroom. Unfortunately, the term “assessment” may mean a variety of different things to different people. For many biologists, assessment is inappropriately simplified to mean anything having to do with grading, exams, or quizzes. Unsurprisingly, scientists who embrace this meaning of assessment have little fondness for it. In the context of college and universities, assessment can additionally bring with it the negative connotations of external evaluation, accreditation, and penalty. Given these preconceptions, it is no wonder that many instructors avoid explorations of assessment at all cost. However, assessment evidence can be a key tool to inform decision making by instructors about what and how to teach the particular students sitting in front of them. From the educational perspective, well-designed assessments are arguably the most effective tools instructors have to foster learning in their classrooms. From a biological perspective, assessment aligns with the very nature of our discipline because assessment is essentially evidence that can inform and drive our decisions about what to do next. Continual, everyday assessment allows classroom instruction—and student learning—to evolve through a structured, iterative process of trial and error that focuses the instructor’s efforts and the students’ energies on those ideas that are the most challenging. We introduce here 13 key terms that biologists would likely encounter in their explorations to learn more about classroom assessment.

12. Summative Assessment and 13. Formative Assessment
A core idea often not understood about assessment is that it can both measure learning and drive learning. “Summative assessment” is the term used to describe assessment activities that measure or evaluate learning (Huba and Freed, 2000). Summative assessments often occur at the end of a learning experience and are often graded, examples being midterm exams and term papers. The grade provides external motivation for performance and acts as an indicator for
external stakeholders to evaluate students. In contrast, “formative assessment” refers to those assessments that occur before or during the learning process and that can be instrumental in driving learning. Formative assessments provide lower-stakes, often nongraded, opportunities for students to check their understanding of a concept or practice thinking through a problem. Examples of formative assessment include minute papers, clicker questions, peer review of written drafts, quick low-stakes quizzes before class, or online tutorials with responsive feedback. Feedback plays a critical role in formative assessment, whether from the instructor, the student’s own self-reflection during the assessment, or discussion of efforts with a peer. Often, the objective of formative assessment is for both instructors and students to evaluate the extent to which students understand. Formative assessments drive learning by providing regular, ongoing feedback to students and instructors about what is being learned, and may or may not be graded.

For further exploration...


As instructors innovate in their teaching approaches, assessment questions offer the opportunity to systematically collect evidence on student learning, fundamentally using our classrooms as learning laboratories. Preassessments and postassessments are simply questions asked of all students before teaching begins and after teaching has concluded. Comparison of preassessments and postassessments can provide instructors evidence about learning gains (or lack thereof), defined as the difference between the beginning and end scores over a designated instructional time period. Postassessments are generally questions posed on summative assessments, such as midterms or final exams. Preassessments offer a baseline about students’ knowledge that can later be compared with postassessments to determine learning gains. More importantly, preassessments—which are woefully uncommon in undergraduate biology classrooms—can also serve other important roles for instructors. Not only can preassessments illuminate what students already know, they can also identify common misconceptions and misunderstandings held by the majority of students. For instructors, this preassessment evidence can be used to tailor the instruction to the particular conceptual challenges of that group of students. When shared with students, preassessment evidence can also focus study efforts toward those ideas. Examples of preassessments can include clicker questions, minute papers, concept maps, or a variety of other student tasks. Additionally, preassessments need not only probe conceptual understanding but can also probe motivation through a questionnaire about why students are taking the class and what they hope to get out of it or probe confidence by having students rank their confidence in their understanding of upcoming concepts. A word of caution: if preassessments are high stakes, especially if connected to grades, they may evoke personal stereotypes among some students and counterproductively impede learning (see 31. Stereotype Threat).

For further exploration...


While assessments can take many forms, they are usually questions, which are either closed-ended questions or open-ended questions. Closed-ended questions provide a finite (closed) set of response options from which the students select what they believe to be the correct answer. Examples of closed-ended questions include multiple-choice, matching, and true–false questions. One benefit of closed-ended questions, such as multiple-choice or matching questions, is that instructors can easily gauge understanding among a population of students (e.g., by using clicker questions). Additionally, closed-ended questions are more quickly graded than open-ended questions. A drawback of closed-ended questions is that—unless carefully crafted—they often assess only lower-order cognitive skills such as remembering terms and are less likely to evaluate more complex skills related to experimentation, synthesis, or communication.

In contrast, open-ended questions provide opportunities for students to compose responses in their own words. Examples of open-ended questions include essay questions, drawing or diagramming questions, or even concept maps. A key benefit of open-ended questions is that they demand students gather and synthesize their ideas and cogently express them in writing. Additionally, open-ended questions are often useful in assessing higher-order cognitive skills related to experimentation, synthesis, or communication. However, open-ended questions do not lend themselves to quick evaluation of understanding among groups of students, nor are they quick to grade. However, strategies such as peer grading or the use of rubrics to facilitate grading make asking open-ended questions tractable. Importantly, open-ended questions may be most useful in driving authentic student learning, since few real-world scientific careers offer closed-ended questions or problems.

For further exploration...


19. Bloom’s Taxonomy

Bloom’s taxonomy—conceived by mid–20th century educator and researcher Benjamin Bloom—is a system with which instructors can judge the nature of the assessment questions
they are asking their students (Bloom et al., 1956). While many have argued about and revised the original categories of Bloom’s taxonomy, the core ideas persist about how to judge the type of thinking that may be elicited by an assessment question. Bloom proposed that learning could be categorized into lower-order cognitive skills (such as knowing, remembering, or describing) and higher-order cognitive skills (such as analyzing, evaluating, inventing, or synthesizing). This simple framework provides instructors with a mechanism to evaluate the extent to which their student learning objectives and their assessment tools are targeting higher-order versus lower-order student thinking.

Importantly, instructors can often judge the nature of their assessment questions based on Bloom’s taxonomy by attending to the verbs used in the assessment. Lower-order assessment questions are often those that ask students to: define, list, describe, explain, summarize, or paraphrase. In contrast, higher-order assessment questions are often those that ask students to: predict, design, apply, defend, propose, or judge. Analysis of the verbs used in either course student-learning outcomes or course assessments with reference to Bloom’s taxonomy can be a useful exercise for instructors who are re-evaluating their approach to teaching.

For further exploration...


20. Classroom Assessment Techniques (CATs)
CATs are a suite of strategies instructors can use to integrate low-stakes, formative assessments into their teaching. The term “CAT” is often used to convey the use of clear, regular, purposeful assessment within a course that drives student learning and provides evidence to guide the learning process for instructors and students. With their influential book Classroom Assessment Techniques, Angelo and Cross (1993) revolutionized the role of assessment for instructors in higher education. The premise of their work is simple: instructors should use regular, low-stakes assessments both during and outside class time to engage students in learning. Many simple CATs were discussed earlier in the section on active learning; examples include minute papers, clicker questions, concept maps, and think–pair–share discussions, to name a few. Depending on the instructional goals, CATs can employ closed-ended assessment questions, open-ended assessment questions, and questions at a variety of Bloom’s taxonomy levels. Importantly, classroom assessment activities can be embedded anywhere in a course: during class, as homework, during discussion sections, as part of online tutorials, or during labs.

For further exploration...


Grading is defined as the point at which the results of instructor assessment activities become evaluations of students’ understanding of the subject. As instructors, grading—in place of learning—can inappropriately appear to be the goal of our teaching efforts. With few exceptions, instructors have to provide a letter grade at the end of the semester for each student. It is a task disliked by many instructors, not only because of the added workload associated with it but also because of the evaluative nature that requires us to make judgments about students’ academic performance. However, instructors have choices about how to approach the translation of assessment evidence into grades. In particular, instructors can choose to construct their course’s learning environment with either criterion-referenced grading or norm-referenced grading. In criterion-referenced grading, the instructor sets a clear set of expectations for achievement and assumes that students who meet those criteria will receive the corresponding grade. For example, a course that uses criterion-referenced grading would include on the syllabus a list of letter grades associated with percent scores in the course, which do not change based on the class performance. An advantage of criterion-referenced grading is that it provides transparency for students about the grading process, so they can self-evaluate their performance. Additionally, collaboration can be encouraged in a criterion-referenced grading environment.

In contrast, norm-referenced grading presumes that final grades will be distributed based on a normal, bell-shaped curve of performance in which students are essentially competing with one another for the top grades. This type of grading—which has a strong history in undergraduate science courses—has several implicit disadvantages: iterative self-evaluation throughout the course becomes less important, collaboration is subtly discouraged, and competition is likely fostered by the instructor, whether intentional or not. As instructors move toward promoting peer instruction, group work, and collaboration through active learning, criterion-referenced grading becomes much more advantageous than norm-referenced grading in constructing active-learning environments.

For further exploration...


24. Rubrics
Rubrics communicate expectations for student performance on a task and can be one of the most powerful tools in an instructor’s toolbox. Rubrics are defined as a set of expectations that illuminate the criteria on which student assignments will be judged. Because they make a set of fixed expectations public, rubrics provide students and instructors with a shared framework to gauge progress and evaluate the work. This is especially important in multifaceted assignments that require critical thinking, content knowledge, and
communications skills. Approaches to developing rubrics vary from holistic to reductionist and from comprehensive to simple. A holistic rubric might describe general expectations for the caliber of work, stating only the criteria for excellence in broad strokes. A more detailed rubric offers multiple categories of criteria, with descriptions of work at different levels, ranging from excellent to inadequate. Whatever the approach, rubrics can afford instructors efficiency in grading students’ work, by making very clear what caliber of work earns what level of points. However, rubrics are arguably even more important for students as a guide in crafting their work assignments, because they are empowered to monitor their own progress against fixed criteria or get feedback from peers and instructors on early drafts of their work.

For further exploration...


EQUITY, DIVERSITY, AND INCLUSIVITY: CREATING FAIR AND ACCESSIBLE LEARNING ENVIRONMENTS

The third pillar of scientific teaching encompasses equity, diversity, and inclusivity. As a biology education community, we focus a great deal of time and energy on issues of “what” students should be learning, as well as on the “how” of teaching. Yet the aspect of classroom teaching that seems to be consistently underappreciated is the nature of “whom” we are teaching. Undergraduate students often appear to be treated as monolithic without attention to research on the pervasive influence that an individual’s personal history and characteristics, culture, and prior experiences in society and in classrooms all have on the teaching and learning processes in our own classrooms. We provide here brief introductions to commonly encountered terms in biology education that are central to understanding, communicating about, and gaining insight into “whom” we are teaching and the efforts underway across the nation to make biology teaching and learning more fair. In the following sections, we introduce 15 key terms that biologists would likely encounter in their explorations of diversity, equity, and inclusivity in biology education.

25. Equity, 26. Diversity, and 27. Inclusivity

All biology education terms, but perhaps especially these—equity, diversity, inclusivity, and inclusive teaching—can have different meanings for different individuals. The spirit of these terms relates to the goal of having all students in a classroom feel overtly included, allowing them to see themselves and their communities in the biology they are learning and to have multiple ways to access and learn the biological ideas. The term “equity” refers to promoting fairness among diverse individuals in a single setting. As an example, only providing access to learning through listening would be deeply unfair to deaf students, which would seem obvious, but also to students who need time to talk through ideas, which may be less obvious. The term “diversity” refers to the myriad differences among students. It tends to conjure ideas about race when used in the United States, and while racial diversity is one key aspect of diversity, so are a number of other axes along which students may see experiences classrooms in different ways: gender, sexual orientation, first-generation college attending, first-generation U.S. citizen, speaking English as a second language, learning disabilities, and the list goes on and on. As such, “inclusivity” and “inclusive teaching” are terms that are intended to encompass all forms of diversity and all attempts at equity into a single idea of creating classrooms where all students feel included.

For further exploration...


28. Student-Deficit Model

“Students just aren’t motivated.” “Nonmajors just aren’t interested in biology.” While each of these statements may seem innocuous, they reflect a student-deficit model approach to explaining challenges in teaching and learning, in which the underlying assumption is that the root of the problem lies with students. Additionally, these statements ignore possible contributions of other factors, such as the teacher (who is often the speaker), the quality of teaching, the structure of the learning environment, or a host of other sources beside students. Student-deficit model thinking is often criticized as unscientific in that it assigns blame to one locus of a complex environment and with little to no evidence. Shifts away from student-deficit models and toward learning environment–deficit models or teaching-deficit models are thought to be key for reforming undergraduate biology education and diversifying the scientific workforce.

For further exploration...


29. Achievement Gap and 30. Instructional Selection

Learning to see inequity in science is critical to anyone who is actively encouraging young people to invest their educations, careers, and lives in the discipline. If the culture of science is grossly inequitable, why should students take the risk of entering this discipline over careers in other arenas? Many scholarly publications from the fields of psychology, science education, and sociology have described inequities in science, proposed theoretical frameworks for understanding them, and explored practical strategies for addressing such inequities (e.g., Tobias, 1990; Seymour and Hewitt, 1997;
Brown, 2004; Johnson, 2007; Tanner, 2007; Chamany et al., 2008), but progress toward eliminating these inequities from our discipline has been slow. By most academic measures, an achievement gap persists from kindergarten through college, defined as a differential between the success of white and Asian students compared with African-American, Latino/a, and Native American students. Some suggest this is due to instructional selection, namely, that traditional, lecture-based learning environments favor students who learn well in that format, while students who prefer other modes are more likely to leave the discipline. Developing an “equity eye”—an ability to analyze the fairness, inequity, and relative participation and success of different types of individuals—causes biologists never to see science classrooms, science conferences, or anything else in their discipline quite the same way ever again.

For further exploration...


31. Stereotype Threat

Stereotype threat is a psychological state in which an individual fears that his or her academic performance might confirm an existing stereotype of a cultural, ethnic, gender, or other group with which he or she is identified. This fear, which may be conscious or subconscious, can then lead to profound impairment of that individual’s academic performance (Steele and Aronson, 1995; Steele, 1999), with some hypothesizing that this impairment is mediated by cortisol stress responses that may impede working memory (reviewed in Schmader et al., 2008). Stereotype threat is thought to most influence those individuals who care deeply about succeeding within the context in which they experience the threat. Stereotype threat can be induced in a variety of ways in experimental psychology laboratory settings, ranging from explicit threats (e.g., stating that women generally do not perform as well on math tests as men) to implicit threats (e.g., having women take a math test in a room in which men are the majority). Importantly, the influence of stereotype threat is profound. As an example, the test performance of women who had been previously shown to perform in the 90th percentile on a challenging math assessment was decreased by almost 50% in a situation of stereotype threat (Johns, 2005). While stereotype threat has been studied to a lesser extent in situ in undergraduate classrooms, ongoing implicit threat situations likely exist in the sciences, given the low proportions of students of color and women in these classrooms. Additionally, it is hypothesized that some undergraduate science instructors may unintentionally use explicit threat language or examples in science classrooms.

For further exploration...

S. Miller and K. D. Tanner

that it usually refers to any student from a nonwhite background, including students from the Middle East and Asian and South Asian countries such as the Philippines, China, India, and Indonesia. Finally, the term “first generation” is often a shortened version of “first-generation, college-attending student,” namely, a student whose parents or other relatives have not had the privilege of access to higher education. Alternatively, the term “first generation” may refer to a student’s citizenship status in the United States, with a first-generation student being the first in his or her family to be born in the United States.

For further exploration...


37. Self-Efficacy

While not a term scientists often encounter, self-efficacy is likely a familiar concept: namely, the perceptions a person holds about his or her ability (or lack thereof) to succeed and achieve. A commonly heard example is a variation on: “I’m just bad at science and math.” Self-efficacy, no doubt, results from a host of influences, including personal background, personality, immediate social interactions, and societal biases and attitudes. However, undergraduate science course instructors have enormous power to influence students’ self-efficacy by their attitudes toward students and comments they make about students and their work in courses. It seems an obvious statement that, if we do not believe in students and promote their positive self-efficacy in science, we risk that those students will not believe in themselves. And if they do not believe in themselves, they have little chance of persisting through the challenges inherent in biology learning.

For further exploration...


38. Unconscious or Implicit Bias

While scientists may like to believe that science is objective and unbiased, researchers in cognitive and social psychology have extensively documented that humans across a variety of professions, including our own, are influenced daily by unconscious bias, which is manifested by treating individuals differently based on their personal characteristics. Unconscious bias may occur when forming opinions of how intelligent students are, how hard they are working, or whether they deserve the benefit of the doubt when grading, for example. Recent studies in university-level science courses suggest that unconscious bias against female students is just as common among female faculty members as male faculty members, suggesting that unconscious bias is a product of societal influences and not strongly correlated with one’s own personal characteristics (Moss-Racusin et al., 2012). Importantly, raising one’s own awareness of unconscious bias and being vigilant in questioning one’s own biases is currently the most immediate action instructors can take to decrease bias in their teaching. For example, tracking (mentally or on paper) the personal characteristics of students (e.g., gender) called on to speak during class discussions can alert an instructor to unconscious tendencies to call preferentially on students of one gender. Additional strategies to decrease unconscious bias include blinding oneself to student identity whenever possible, especially when grading exams, which is easily done by having students write their names on the back of exams.

For further exploration...


39. Equitable Teaching Strategies

Designing learning environments that attend to individual students and their interactions with one another may seem an impossible task in a course of 20 students, much less a course of more than 700. However, there are a host of simple teaching strategies, rooted in research on teaching and learning, that can support biology instructors in making classrooms fair and inclusive. These teaching strategies are sometimes referred to as “equitable teaching strategies,” wherein striving for classroom equity is about teaching all the students in your classroom not just those who are already engaged, already participating, and perhaps already know the biology being taught. Equity, then, is about striving to structure biology classroom environments that maximize fairness, environments in which all students have opportunities to verbally participate, all students can see their personal connections to biology, all students have the time to think, all students can pose ideas and construct their knowledge of biology, and all students are explicitly welcomed into the intellectual discussion of biology. Without attention to the structure of classroom interactions, what can often ensue is a biology lesson that can be accessed by only a small subset of students in a classroom. Examples of equitable teaching strategies include simple ways instructors can actively structure classroom interactions such as: 1) think–pair–share, 2) wait time, 3) multiple hands, multiple voices, 4) allowing students time to write, as well as more involved strategies such as 5) integrating culturally diverse examples, 6) using varied active-learning strategies, 7) being explicit about promoting access and equity, and 8) establishing classroom community and norms (Tanner, 2013).

For further exploration...


MOVING TO PRACTICE: INSTRUCTIONAL DESIGN, LEARNING, AND TECHNOLOGIES

While the ideas of scientific teaching—active learning, assessment, and inclusivity—may intrigue many instructors, translating these ideas into classroom teaching practices can prove challenging for even experienced instructors. In particular, incorporating scientific teaching into practice requires instructors to refocus their efforts on student learning rather than teaching. To enact this philosophical shift at the practical level, instructors can use a variety of tools to make small changes to their teaching that bring together active learning, assessment, and equity considerations, described in the previous sections. We introduce here 11 key terms that relate to tools and strategies to support biologists in translating scientific teaching into practice.

40. Backward Design

In their influential book Understanding by Design Wiggins and McTighe (1998) rejected the traditional approach to course design, which generally begins with writing a syllabus that follows the order of the textbook and, later, writing exams after having delivered a series of lectures. Instead, they proposed that instructors could best support student learning by approaching course design “backward,” namely, by determining what students should be able to do at the end of a course, designing the assessments by which this would be measured, and then working backward to design learning experiences to get them there. Wiggins and McTighe’s approach was termed “backward design,” because the instructor begins planning with the end goals of the course, namely, student learning outcomes and the assessments by which they will be evaluated, then designs the learning experiences. The approach is simple. First, instructors determine what they aspire for students to be able to understand and do with that understanding by the end of a course, establishing clearly these student learning outcomes. Then, instructors engage in designing assessment tools that can be used to gauge student understanding early in the course and to evaluate the extent to which learning outcomes are achieved. Only at this point would instructors begin to craft learning experiences—including readings, homework problems, projects, and/or in-class activities—that would support students in their learning and in achieving the instructors’ projected student learning outcomes. While backward design—much like the scientific method—is in theory linear, it is in practice an iterative, nuanced process that requires constant reevaluation of the alignment among learning outcomes, assessments, and learning activities.

For further exploration...


41. Learning Theory and 42. The 5E Model

Learning theory proposes that students should have ample opportunities to interact with the material they are learning. Many researchers have proposed variations on learning models, but most share a few common elements, captured well by Ambrose et al. (2010). First, students’ prior knowledge can positively or negatively influence learning. Second, how students organize the knowledge they are acquiring influences both how they learn and how they will be able to apply what they know. Third, students’ motivation determines, directs, and sustains what they do to learn. Fourth, for students to develop mastery, they must acquire skills, practice them, and know when to apply them. Fifth, goal-directed practice, accompanied by targeted feedback, enhances the quality of learning. Sixth, students’ own levels of development interact with the social, emotional, and intellectual climate of the classroom to influence the learning process. Seventh, to become self-directed learners, students must learn to monitor and adjust their approaches to learning (see 43. Metacognition). While these tenets of learning theory seem like important considerations, instructors often struggle with how to account for these ideas in planning learning activities for students.

The 5E model (Bybee et al., 2006), translates what is known about how humans learn from research in various disciplines—cognitive science, psychology, and science education—into a tool that guides instructors in planning effective learning experiences for students. The five “E”s—engage, explore, explain, elaborate, and evaluate—capture the optimal order in which students navigate learning. First, students must be interested and engaged in the concepts at hand. Then, learning is best facilitated when students have opportunities to explore their prior knowledge of the topic. At that point, instructors and students are likely ready to engage in coexamination and coexplanation of the ideas under study. At this point, students need opportunities for deliberate practice through elaboration—application of the concepts learned to new contexts or problems. And finally, upon completion of these components of the cycle, students and instructors (together) evaluate what students have learned. The 5E model may be a helpful tool instructors can use to analyze their current approaches to teaching, identifying which aspects (“E”s) of the learning cycle may be missing from their instruction and which pieces of their current instruction can simply be reordered to provide students with more opportunities for engagement and exploration before explanations are offered.

For further exploration...

43. Metacognition

To become self-directed learners, students must learn to develop a “meta-level” awareness of their learning through which they can monitor and adjust their approaches to learning; this is called “metacognition.” During this process, students self-evaluate what they are and are not understanding in a course and determine what actions might set them on a better path to achieving the learning outcomes. Metacognition can be approached in a variety of ways. For example, instructors can use preassessments to encourage students to examine their current thinking about biological concepts before considering any new instruction on the concept at hand; students are then empowered to focus on what they need to learn and to develop strategies to do so. Additionally, instructors can ask students outright to identify their confusions through in-class minute papers or homework assignments; again, raising awareness of where the students need to learn more helps prioritize what and how to study. Perhaps most importantly, metacognition can be encouraged in biology classrooms by giving students the license to be incorrect and share their nascent thinking, which may be scientifically inaccurate, as part of the learning process. Teaching students to use metacognition to understand how they are thinking about biology provides an important step on the path to thinking like a biologist.

For further exploration...


44. Syllabus

The syllabus is a document that offers the first opportunity for faculty members to communicate with their students about the course and provides a guiding framework that sets the tone and expectations for the semester. The conventional syllabus is composed of a calendar that relays the sequential order of topics to be covered during the semester, along with other information about the course meeting times and contact information. A more detailed, arguably learner-focused syllabus might also describe what learning outcomes students should achieve by the end of the course, outline how learning will be assessed, or delineate the responsibilities of both instructor and students for achieving the course goals. In the syllabus, instructors have the opportunity to set the tone of the learning environment they hope to cultivate, minimize student resistance, and focus students on strategies for learning.

For further exploration...


45. Lesson Plans

Lesson plans are outcome-oriented organizers that provide the day-by-day details and supporting materials for an instructional plan. For some instructors who currently keep few records related to their teaching beyond lecture slides and notes, lesson plans are a first step toward more purposefully planned class time. How? A well-crafted lesson plan allows the course outcomes to manifest as real actions by linking outcomes to activities (what the instructor and students are both doing), connecting them with the assessments (what is being gauged or graded at any point), and noting the actual times these things should happen (in and out of class). They include detailed instructor notes that explain how to launch and wrap up student activities, show how to transition to a technology-enabled portion of the class period, predict what students will say (e.g., questions students might ask or brainstorm comments they might make, plus notes about how to deal with them), aggregate any notes and PowerPoint slides, provide any student handouts or materials (e.g., clickers, note cards, etc.), and collate any other preparatory materials the instructors or students might need. In addition, they can serve as the first step toward collecting evidence about the effectiveness of the instruction, enabling iterative improvement in future offerings of the course.

For further exploration...

46. Learning Technologies, 47. Blended Learning, and 48. Universal Design

Learning technologies encompass all the instructional technologies that can be integrated into college courses across a spectrum of formats, ranging from fully face-to-face meetings to completely online environments. Most courses naturally fall somewhere in between, in a space dubbed “blended learning,” in which some aspects of teaching and learning occur face-to-face and others online. This term captures the integrated spirit and structure of these courses, which “are instructor-designed and supervised environments that use face-to-face and technology-mediated channels to enhance interactive, engaging learning experiences and to improve student learning outcomes” (https://blendedtoolkit.wisc.edu). When integrating technology, it is critical to consider not just the learning outcomes and the teaching approaches used to achieve these outcomes, but also the managerial, technical, and social aspects of the instruction. Finally, technology elevates issues related to accessibility. Principles of universal design ensure that online and computer-aided elements are accessible to all audiences by accounting for particular disabilities, such as visual or auditory impairments, or varied computer platforms.

For further exploration...

49. Scientific Teaching and 50. Nature of Science in the Classroom

The crux of scientific teaching is twofold: the instructor approaches teaching with the same guiding practices and philosophies of scientific research, while in parallel affording opportunities for students to experience the iterative process of thinking and discovery that is characteristic of science. Instructor attention to the tenets of scientific teaching—active learning; assessment; and equity, diversity, and inclusion—are the means by which instructors can construct effective learning environments for students. From the instructor perspective, it is an intentional approach to teaching that involves questioning, evidence, and innovation. In the same spirit, students learn how to engage in asking questions, using evidence to answer those questions, and to be creative in solving real-world problems; this is the nature of science. And the more students and instructors integrate scientific practices and habits of mind into everything about the teaching and learning of college biology, the more authentic each classroom experience will be, and the more each student will experience being a scientist. In this paper, we have unpacked some of the key components of scientific teaching that were originally explored in the book Scientific Teaching (Handelsman et al., 2006) and highlighted the key terms and practices we believe are most critical for the new or future instructor.

For further exploration...


IN CONCLUSION

We have attempted here to offer readers at various stages of their careers a portal into the language of biology education. Whether you are a graduate student or postdoctoral fellow, a recently hired faculty member teaching for the first time or a seasoned classroom veteran seeking new ideas and research literatures in biology education, we encourage you to continue your explorations of biology education far beyond these introductory paragraphs. It is an incredibly exciting time to be engaged in biology education efforts, as the disciplines of neurobiology, science education, psychology, and cognitive science increasingly overlap in the focus of their research. No doubt, just like the basic biological mechanisms that we study, these biology education ideas will change over time with new biology education research and emerging evidence, and we hope that you will follow these developments with as much enthusiasm as you follow new discoveries in other fields.

REFERENCES


Rowe MB (1987). Wait time: slowing down may be a way of speeding up. Am Educator 11, 38–43, 47.


Tanner KD (2009b). Talking to learn: why biology students should be talking in classrooms and how to make it happen. CBE Life Sci Educ 8, 89–94.


Feature
WWW.Life Sciences Education

Online Resources for Understanding Outbreaks and Infectious Diseases
Nicola C. Barber and Louisa A. Stark

Genetic Science Learning Center, University of Utah, Salt Lake City, UT 84108

Disease outbreaks are hot topics that often receive extensive national and international news coverage, although this coverage may not always be accurate. Engaging students with these current events can be a powerful way to teach about science and health. Accurate disease information is also an important public health issue, as misinformation can lead to fear and poor policy decisions. In this review, we highlight online resources for teaching about outbreaks and infectious diseases that will be useful for scientists and educators working with middle school, high school, and undergraduate students. We particularly focus on current news about infectious diseases, epidemiology, pathogen biology, and vaccines.

Infectious diseases are caused by pathogenic microbes, including viruses, bacteria, and parasites. Some diseases, such as the flu, spread from human to human, while others rely on a vector intermediate—for example, malaria is transmitted by mosquitoes. Outbreaks, or epidemics, are characterized by the rapid spread of disease in a population. An outbreak that becomes global is referred to as a pandemic. In 2009, the global spread of H1N1 “swine” flu was a pandemic, whereas the outbreak of Ebola that began in 2014 is an epidemic in West Africa, not a pandemic.

OUTBREAK NEWS AND INFORMATION
Several websites provide extensive information on infectious diseases and outbreaks. The World Health Organization (WHO) is an international public health agency within the United Nations and a global authority on disease outbreaks, epidemiology, and health. Its website (www.who.int) provides news and is a reliable source for the latest data on global health and disease. The Global Alert and Response (GAR) program is particularly relevant for infectious diseases (www.who.int/csr/en). Disease-specific data and information can be found in the Diseases section, located in the left menu. GAR also runs the Disease Outbreak News page (www.who.int/csr/don/en). These updates from around the globe include specific cases, public health response, risk assessment, and advice; updates may be sorted by year, disease, or country. The WHO website offers a wealth of information, but it can be somewhat difficult to navigate. Fortunately, current outbreaks are often linked directly from the home page, making navigation to outbreak information easier.

The Centers for Disease Control and Prevention (CDC) is the national public health agency of the United States. Its website (www.cdc.gov) maintains up-to-date information about disease and health both for the United States and abroad. The Outbreaks section of the CDC home page lists recent outbreaks and links to the CDC Current Outbreak List (www.cdc.gov/outbreaks/index.html). Outbreaks are sorted by location (U.S.-based and international) with links to information about each one. Information on specific diseases also can be found using the Diseases & Conditions pull-down menu with links to information about each disease. Content about current outbreaks is updated frequently, and the website is easy to navigate, despite containing a large amount of information and resources.

Bridging the gap between public health agencies and mainstream news media, the University of Minnesota’s Center for Infectious Disease Research and Policy (CIDRAP) website (www.cidrap.umn.edu) is an excellent source for...
the latest news about infectious diseases. Led by a team of public health professionals, CIDRAP combines expertise in research, policy, education, and news media. The CIDRAP home page is an easy-to-navigate online newspaper with original content on emerging diseases and outbreaks and is updated every weekday (Figure 1). Disease-specific content is accessed through the Infectious Disease Topics list in the top menu bar and includes more than 60 topics ranging from specific diseases such as norovirus to more general topics such as childhood vaccines and bioterrorism. These links provide easy access to detailed content and include menu tabs for Recent News, Resources, and Literature. The Literature tab includes a list of recent publications and time-stamped links to information and resources from national and international health agencies, professional organizations, and academic journals. In addition, many of the diseases, such as avian influenza and anthrax, have comprehensive overviews that synthesize information about each disease’s biology, epidemiology, clinical features, treatment, and current public health considerations and recommendations. The Public Health Practices section provides links to resources, including fact sheets, training materials, posters, and K–12 educational materials. The greatest strengths of this website are its ease of navigation, embedded links, lists of relevant research literature, and up-to-date content.

EPIDEMIOLOGY AND DISEASE DETECTIVES

Epidemiology is the study of the incidence, spread, cause, and effect of disease in a population. The Marian Koshland Science Museum of the National Academy of Sciences has an online exhibit called Infectious Disease: Evolving Challenge to Human Health (www.koshland-science-museum.org/explore-the-science/interactives/infectious-disease#.VHOgilF-K8). This virtual exhibit uses interactives to provide an introduction to epidemiology, disease biology, and pathogen evolution. The Virtual Field Trips section of the educator resources provides lessons using guided-inquiry approaches to exploring the online materials for middle and high school students. The middle school materials also include a classroom activity for simulating the spread of an infectious disease and tracing its source.

The cause of a disease is often mysterious. The CDC’s Solve the Outbreak game (Figure 2) allows the user to play the role of an epidemiologist, or disease detective, and is...
Outbreaks and Infectious Diseases

available online at www.cdc.gov/mobile/applications/sto/web-app.html and through the Apple and Android App stores. Users progress through story-based levels, answering questions in response to the clues and data provided. Although users are often given choices of what to do next, they will not advance in the game until they select the best answer. A How to Play tab is available, but most users will find the sleek interface intuitive. There is not much replay value once a case is solved, but a perfect score on the first 12 outbreaks in level 1 is required to access four new outbreaks in level 2. Throughout, the game provides links to relevant content on the CDC website, including details about real outbreaks. A Learn tab provides information on epidemiology and disease detectives, the Epi Curve tool, and careers in epidemiology and laboratory science. Badges are awarded for both game performance and accessing certain information.

Rice University’s MedMyst (http://medmyst.rice.edu) also uses a disease detective game play to teach about the spread of disease. MedMyst is available in English and Spanish and is aimed at middle school students. The two most recent MedMyst Reloaded games are Disease Defenders and Animal Alert! (Figure 3). Both games engage users in investigating an infectious disease outbreak using science process skills and learning about the roles of epidemiologists, microbiologists, and veterinarians in an outbreak response. Game play is largely story line based and punctuated with more game-like activities, some of which are more closely tied to learning than others. A strength of MedMyst is that it portrays career possibilities in an inclusive and accessible way. The MedMyst educator site includes worksheets, classroom activities, and assessment items, and ties the games and materials to science standards.

PATHOGEN BIOLOGY

For delving into the biology of the pathogens themselves, a good starting point is the Microbe World website (www.microbeworld.org/index.php) produced by the American Society for Microbiology. The home page includes microbiology news and research updates, many of which focus on microbes that are not pathogens. Although some pathogenic microbes cause devastating diseases, it is important for students to learn that there are many more benign or beneficial microbes. Content on the website is largely text based with accompanying images or illustrations. The most recent videos and images can be found through tabs at the top of the site’s home page. To find information on a specific topic, use the keyword search in the top right corner; the search can be filtered by content type. Of particular relevance to infectious disease biology is a behind-the-scenes video (56:48 min) of researchers entering a biosafety level 4 laboratory where they must wear airtight suits to study the most dangerous infectious disease agents, such as the Ebola virus and the plague-causing bacteria Yersinia pestis (www.microbeworld.org/component/content/article?id=1362). Much of the information on the site is most appropriate for undergraduate students. At the bottom of the page, middle and high school students can find basic information about microbes and microbiology that would be appropriate for them.
To provide engaging and accessible information about vaccine science and public health, the College of Physicians of Philadelphia has created an interactive website called The History of Vaccines (www.historyofvaccines.org). This website is easy to navigate, with tabs for timelines of vaccine history, activities, articles, and an image gallery. The Activities section contains interactives on how vaccines work, types of vaccines, and how vaccines are made. It also includes the Illsville game (Figure 4), which combines the history of medicine with concepts about epidemiology, quarantine, and vaccine development.

**VACCINES**

Vaccination is a powerful means of combating infectious disease. There is significant interest in vaccines but also considerable misinformation about them. People are often curious about what vaccines are, how they work, and why we do not have a vaccine for every disease. Both WHO and the CDC provide some basic information about vaccines from the public health perspective at www.who.int/influenza/vaccines/en and www.cdc.gov/vaccines, respectively.

**Figure 3.** MedMyst’s Animal Alert! game allows users to choose a career path to investigate a disease outbreak (http://webadventures.rice.edu/ed/Teacher-Resources/_games/MedMyst-Reloaded/_402/Game-Overview.html).

**Figure 4.** The History of Vaccine’s Illsville game teaches the history of medicine, explains how infectious diseases spread, and gives an overview of vaccines (www.historyofvaccines.org/content/illsville-fight-disease).
vaccination, and herd immunity. Players respond to disease outbreaks in a population by employing quarantine, doctors, and educators. The game ends with a vaccinated population that has achieved herd immunity. The Articles section of the website contains more details, including information on the challenges of vaccine design, why some diseases are difficult to target with vaccines, viral evolution, and optimistic news about vaccines against Ebola and HIV. The Educators link on the top right of the navigation bar links to the online activities and associated lesson plans. The website is available in English and Spanish.

IMAGES

Images of pathogens and public health efforts will be useful for anyone doing outreach presentations. The CDC’s Public Health Image Library (http://phil.cdc.gov/phil/home.asp) is a searchable database containing health-related images, including photos of public health efforts and electron micrographs of microbes. The influenza section includes several three-dimensional graphical representations of a generic influenza virion’s ultrastructure. Clicking on an image provides a description and additional information. Graphics and images from the National Institute of Allergy and Infectious Disease (NIAID) include a striking image of an Ebola-infected cell (Figure 5). Many images can be found in their public Flickr account at www.flickr.com/photos/niaid.

ADDITIONAL RESOURCES FOR PRECOLLEGE TEACHERS

For middle and high school teachers, several epidemiology modules with detailed lesson plans are available online. The National Institutes of Health Curriculum Supplement Series on Emerging and Re-emerging Infectious Diseases tackles infectious diseases, epidemiology, public health, and policy at http://science.education.nih.gov/supplements/nih1/Diseases/default.htm. Published in 1999, these lessons include several online videos and contain five detailed activities targeted for the high school level. Montclair State University’s Detectives in the Classroom epidemiology curriculum (www.montclair.edu/Detectives/crriculum/ CurriculumModules.htm) contains five modules aimed at middle and high school students; the six lessons in module 1 are relevant to infectious diseases. The CDC’s BAM! Body and Mind website (www.cdc.gov/bam) offers text and illustrations aimed at ages 9–13. It covers a variety of topics, including pathogens, the immune system, and vaccines. The corresponding teacher site (www.cdc.gov/bam/teachers/index.html) has an infectious disease epidemiology module with classroom lessons that introduce microbes, epidemiology, and the scientific process to younger students.

People are naturally curious about health and disease. Outbreaks and infectious diseases can be a powerful way to integrate current events, health, medicine, and science in the classroom. It is also beneficial to our communities when we share what we know to better inform students and policy makers. The websites reviewed here will make your task easier and can make a difference in the lives of students looking for careers that will be meaningful to them.

ACKNOWLEDGMENTS

We thank A. Malcolm Campbell for critical comments and helpful suggestions on this Feature.
Feature

Current Insights

Recent Research in Science Teaching and Learning

Deborah Allen

Department of Biological Sciences and Center for Teaching & Assessment of Learning, University of Delaware, Newark, DE 19716

This feature is designed to point CBE—Life Sciences Education readers to current articles of interest in life sciences education as well as more general and noteworthy publications in education research. URLs are provided for the abstracts or full text of articles. For articles listed as “Abstract available,” full text may be accessible at the indicated URL for readers whose institutions subscribe to the corresponding journal.


Online publication of this meta-analysis last spring no doubt launched a legion of local and national conversations about how science is best taught—as the authors state the essential issue, “Should we ask or should we tell?” To assess the relative effectiveness of active-learning (asking) versus lecture-based (telling) methods in college-level science, technology, engineering, and mathematics (STEM) classes, the authors scoured the published and unpublished literature for studies that performed a side-by-side comparison of the two general types of methods. Using five predetermined criteria for admission to the study (described fully in the materials and methods section), at least two independent coders examined each potentially eligible paper to winnow down the number of eligible studies from 642 to 225. The working definition of what constitutes active learning (used to determine potential eligibility) was obtained from distilling definitions written by 338 seminar attendees; what constitutes lecture was defined as “continuous exposition by the teacher” (quoted from Bligh, 2000). The eligible studies were situated in introductory and upper-division courses from a full range of enrollment sizes and multiple STEM disciplines and included majors and nonmajors as participants. The frequency of use and types of active-learning methodologies described in the 225 eligible studies varied widely.

Quantitative analysis of the eligible studies focused on comparison of two outcome variables: 1) scores on identical or formally equivalent examinations and 2) failure rates (receipt of a “D” or “F” grade or withdrawal from the course). Major findings were that student performance on exams and other assessments (such as concept inventories) was nearly half an SD higher in active-learning versus lecture courses, with an effect size (standardized mean weighted difference) of 0.47. Analyses also revealed that average failure rates were 55% higher for students in the lecture courses than in courses with active learning. Heterogeneity analyses indicated that 1) there were no statistically significant differences in outcomes with respect to disciplines; 2) effect sizes were lower when instructor-generated exams were used versus concept inventories with both types of courses (perhaps because concept inventories tend to require more higher-order thinking skills); 3) effect sizes were not significantly different in nonmajors versus majors courses or in lower versus upper-division courses; and 4) although active learning had the greatest positive effect in smaller-enrollment courses, effect sizes were higher with active learning at all enrollment sizes. Two types of analyses, calculation of fail-safe numbers and funnel plots, supported a lack of publication bias (tendency to not publish studies with low effect sizes). Finally, the authors demonstrated that there were no statistically significant differences in effect sizes despite variation in the quality of the controls on instructor and student equivalence, supporting the important conclusion that the differences in effectiveness between the two methods were not instructor dependent.

In one of the more compelling sections of this meta-analysis, the authors translated the relatively dry numbers resulting from statistical comparisons to potential impacts on the lives of the students taking STEM courses. For example, for the 29,300 students reported for the lecture treatments across all students, the average difference in failure rates (21.8% in active learning vs. 33.8% with lecture) suggests that 3516 fewer students would have failed if enrolled in an active-learning course. This and other implications for the more beneficial impact of active learning on STEM students led the authors to state, “If the experiments analyzed here had been conducted as randomized controlled trials of medical interventions, they may have been stopped for benefit.” That is, the control group condition would have been halted...
because of the clear, beneficial effects of the treatment. The authors conclude by suggesting additional important implications for future undergraduate STEM education research. It may no longer be justified to conduct more “first-generation” research comparing active-learning approaches with traditional lecture; rather, for greater impact on course design, second-generation researchers should focus on what types and intensities of exposure to active learning are most effective for different students, instructors, and topics.


This provocative commentary by Carl Weiman highlights the major findings reported in the Proceedings of the National Academy of Sciences by Freeman et al. (2014) and underscores the implications. The graphical representations displaying the key data on effect sizes and failure rates presented in the Freeman et al. meta-analysis are redrawn in the commentary in a way that is likely to be more familiar to the typical reader, making the differences in outcomes for active learning versus lecture appear more striking. Weiman concludes by elaborating on the important implications of the meta-analysis for college-level STEM educators and administrators, suggesting that it “makes a powerful case that any college or university that is teaching its STEM courses by traditional lectures is providing an inferior education to its students. One hopes that it will inspire administrators to start paying attention to the teaching methods used in their classrooms ... establishing accountability for using active-learning methods.”


National societies, committee reports, and accrediting bodies recommend that engineering curricula be designed to prepare future engineers for the complex interdisciplinary nature of the field and for the multitude of skills and perspectives they will need to be successful practitioners. The authors posit that case-based instruction, with its emphasis on honing skills in solving authentic, interdisciplinary, and ill-defined problems, aligns well with these recommendations. However, the methodology is still relatively underutilized, and its effectiveness is underexamined. This article describes a study designed to advance these issues by comparing lecture- and case-based methods within the same offering of a 72-student, upper-level, required course in mechanical engineering.

The study used a within-subjects, posttest only, A-B-A-B research design across four key course topics. That is, two lecture-based modules (the A or baseline phases) alternated with case-based modules (the B or treatment phases). Following each module, students responded to open-response quiz questions and a survey about learning and engagement (adapted from the Student Assessment of Learning Gains instrument). The quiz questions assessed ability to apply knowledge to problem solving (so-called “traditional” questions) and ability to explain the concepts that were used (“conceptual” questions). This study design had the advantage that the same students experienced both the baseline and treatment conditions twice. The authors describe in detail the pedagogical approaches used in both sets of the A and B phases.

The quizzes were scored by independent raters (with high interrater reliability) on a 0–3 scale; scores were analyzed using appropriate statistical methods. Survey items were analyzed using a principal-components factor analysis; composite scores were generated for a learning confidence factor and an engagement–connections factor. Analyses revealed that the two pedagogical approaches had similar outcomes with respect to the traditional questions, but conceptual understanding scores (indicating better understanding of the concepts that were applied to problem solving) were significantly higher for the case-based modules. Students reported that they appreciated how cases were better than lecture in helping them make connections to real-world concerns and see the relevance of what they were learning, but there were no significant differences in students’ perceptions of their learning gains in the case-based versus the lecture modules. The authors note that many studies have likewise demonstrated that students’ perceptions of their learning gains in more learner-centered courses are often not accurate reflections of the actual learning outcomes.

The authors conclude that while these results are promising indications of the effectiveness of case-based instruction in engineering curricula, the studies need to be replicated across a number of semesters and in different engineering disciplines and extended to assess the long-term effect of case-based instruction on students’ ability to remember and apply their knowledge.

Although this study was limited to an engineering context, the case-based methodologies and research design seem well-suited for use in action research in other disciplines.


Well-documented challenges to conceptual change faced by students of evolution include the necessity of unseating existing naïve theories (such as natural selection having purposiveness), having the ability to view the complex and emergent nature of evolutionary processes through systems-type thinking, and being able to see the connections between evolutionary content learned in the classroom and everyday life events that can facilitate appreciation of its importance and motivate learning. To help students meet these challenges, the authors adapted a pedagogical model called Teaching for Transformative Experiences in Science (TTES) in the course of instruction on six major concepts in evolutionary biology. This article reports on a comparison of the effectiveness of TTES approaches in fostering conceptual change and positive affect with that of instruction enhanced with use of refutational texts (RT). Use of RTs to promote conceptual change, a strategy with documented effectiveness, entails first stating a misconception (the term used by the authors), then explicitly refuting it by elaborating on a
By contrast, the TTES model promotes teaching that fosters transformative learning experiences—teaching in which instructors 1) place the content in a context allows the students to see its utility or experiential value; 2) model their own transformative experiences in learning course concepts; and 3) scaffold a process that allows students to rethink or “reeze” a concept from the perspective of their previous, related life experiences.

The authors designed the study to address three questions relevant to the comparison of the two approaches: would the TTES group (vs. the RT group) demonstrate or report 1) greater conceptual change, 2) higher levels of transformative experience, and 3) differences in topic emotions (more positive affect) related to learning about evolution? The study used three survey instruments, one that measured the types and depth of students’ transformative experiences (the Transformative Experience Survey, adapted from Pugh et al., 2010), another that assessed conceptual knowledge (Evolutionary Reasoning Scale; Shulman, 2006), and a third that evaluated the emotional reactions of students to the evolution content they were learning (Evolution Emotions Survey, derived from Broughton et al., 2011). In addition to Likert-scale items, the Transformative Experience Survey contained three open-ended response questions; the responses were scored by two independent raters using a coding scheme for degree of out-of-school engagement. The authors provide additional detail about the nuances of what these instruments were designed to measure and their scoring schemes and include the instruments in the appendices. The Evolutionary Reasoning Scale and the Evolution Emotions survey were administered as both pre- and posttests, and the Transformative Experience survey was administered only at the end of the intervention. The treatment (TTES, n = 28) and comparison (RT, n = 27) groups were not significantly different with respect to all measured demographic variables and the number of high school or college-level science courses taken.

Briefly, the evolutionary biology learning experience that participants were exposed to was 3 d in duration for both the treatment and comparison groups. On day 1, the instructor (the same person for both groups) gave a PowerPoint lecture on the same six evolutionary concepts, with illustrative examples. For the treatment group only, the instructor drew from his own transformative experiences in connection with the illustrative examples, describing how he used the concepts, what their value was to him, and how each had expanded his understanding and perception of evolution. On days 2 and 3 for the treatment group, the students and instructor engaged in whole-class discussions about their everyday experiences with evolution concepts (and related misconceptions) and their usefulness; the instructor scaffolded various “reezing” experiences throughout the discussions. For the comparison group, misconceptions and refutations were addressed in the course of the day 1 lecture, and on days 2 and 3, the participants read refutational texts and then took part in discussions of the texts led by the instructor.

Survey results and accompanying statistical analyses indicated that both groups exhibited gains (with significant t statistics) in understanding of the evolution concepts as measured by the Evolutionary Reasoning Scale (Shulman, 2006). However, the gains were greater for the treatment (TTES) group: effect size, reported as a value for eta-squared, η², equaled 0.29. The authors point out by way of context for this outcome that use of RTs, along with follow-up discussions that contrast misconceptions with scientific explanations, has been previously shown to be effective in promoting conceptual change; thus, the comparison was with a well-regarded methodology. Additionally, the Transformation Experience Survey findings indicated higher levels of transformative experience for the TTES group participants; they more extensively reported that the concepts had everyday value and meaning and expanded their perspectives. The TTES group alone showed pre- to posttest gains in enjoyment while learning about evolution, a positive emotion that may have classroom implications in terms of receptivity to learning about evolution and willingness to continue study in this and related fields.

The authors conclude that the TTES model can effectively engage students in transformative experiences in ways that can facilitate conceptual change in content areas in which that change is difficult to achieve. In discussing possible limitations of the study, they note in particular that the predominance of female study participants (71% of the total) argues for its replication with a more diverse sample.

I invite readers to suggest current themes or articles of interest in life sciences education, as well as influential papers published in the more distant past or in the broader field of education research, to be featured in *Current Insights*. Please send any suggestions to Deborah Allen (deallen@udel.edu).

**REFERENCES**


Feature

Book Review

**Integrating Concepts in Biology: A Model for More Effective Ways to Introduce Students to Biology**


K. N. Prestwich and A. M. Sheehy

Department of Biology, College of the Holy Cross, Worcester, MA 01610

*Integrating Concepts in Biology* (ICB) is the apt title of this groundbreaking electronic textbook (see Supplemental Material). The target audience is students seeking an introduction to biology. It is structured to focus student attention on key concepts underlying biology at all levels of organization. In contrast to the current encyclopedic model of an introductory textbook, this e-book makes effective use of electronically linked text and online resources, selects examples to explore in depth, and offers students means to deepen their understanding. It explicitly uses current ideas about student learning, supported by evidence of best practices, to help students develop as biologically literate thinkers. The authors are colleagues at Davidson College: biologists A. Malcolm Campbell (cellular and molecular biology) and Christopher J. Paradise (ecology) and mathematician/computer scientist/bioinformaticist Laurie J. Heyer. They share a common passion for innovative pedagogy and multidisciplinarity, and ICB is their laudable effort to address many of the widely recognized problems evident in introductory biology texts and introductory courses and explored in recent reports such as *BIO2010* (National Research Council [NRC], 2003) and *Vision and Change in Undergraduate Biology Education* (American Association for the Advancement of Science, 2011).

**THE TEXT**

ICB is organized to emphasize the unity of biological science across the size/complexity hierarchy. It focuses students on five unifying themes (“big ideas”): information, evolution, cells, homeostasis, and emergent properties. The text is divided into 10 major sections, five of which center on
each of the “big ideas” at the molecular, cellular, and organ
system levels, with the other five examining the “big ideas”
at organismal through ecosystem levels. This division is
designed to allow ICB to be used in a typical two-semester
introductory sequence. Each of the 10 sections consists of three
chapters. For example, the grouping Information in Mole-
cules and Cells contains chapters 1–3, respectively titled
“Heritable Material” (molecules), “Central Dogma” (cells),
and “Reproduction and Cell Division” (cells and organ-
isms). An overview of the text’s structure is encapsulated in
a graphic shown at the top of the main page for each chapter
(Figure 1).

Each of ICB’s 30 chapters is subdivided into approximately
three sections of text focused on an interesting question
(“Can non-living objects compete and grow?”) that often has
implications beyond biology (“Why is [the herbicide] para-
quiat used in America but illegal in Europe?”). Chapter sec-
tions contain about the right amount of material to form the
nucleus of one class period. Content is focused on important
examples of what the authors consider to be illustrations of
the five big ideas around which the text is organized. The
examples are carefully chosen, and they are expounded in a
way that encourages students to engage data using the same
processes as a practicing biologist. There is far less emphasis
on new vocabulary and often less detail than would be seen
in a more traditional text. For example, complicated path-
ways (e.g., glycolysis, Krebs cycle, mitochondrial electron
transport, photosynthesis) are reduced to their most salient
features (few individual reactions are considered). Accom-
panying this “bare-bones” description are some of the data
used by the two Krebs and their coworkers to discover that
the mitochondrial citric acid pathway was actually a cycle.
Students are guided through the data and encouraged to
add to what they were told about metabolism and develop
a sense of how biochemists come to understand the oper-
ation of complex pathways. Likewise, the general function
of ATP synthase is sketched out, but it is accompanied by a
series of questions and online resources that help students
understand how the structure and function of this complex
molecule have been (and are being) worked out. At various
points in each section, students are prompted to make con-
nections and build on what they learned previously (e.g., ap-
ply knowledge of the mitochondrial electron transport sys-
tem and chemiosmosis to interpretation of data dealing with
chloroplasts and the light reactions). At the end of each sec-
tion, review questions are presented. The goal is to provide
gentle guidance and inspiration as students construct their
own knowledge base and develop their analytical skills.

Nearly all chapters include one or more Bio-Math Explora-
tions (BMEs), and all have at least one Ethical, Legal, Social
Implications (ELSI) section. These are well integrated with
the rest of the text. The BMEs introduce students to quantifi-
cation, the uses of statistics to summarize and interpret data,
and mathematical models and other tools. They are often
accompanied with spreadsheets for making repetitive calcu-
lations and/or they make use of linked simulations. BMEs
vary in difficulty from easily manageable (e.g., a worksheet
calculation of a Shannon [ecological] diversity index) to
challenging (e.g., “How can you quantify a pattern in a gene
sequence?”), and they are unusually sophisticated for an in-
troductory course (e.g., “How can you count animals you
cannot see?” goes beyond the commonly taught Lincoln in-
dex). Likewise, the ELSI sections are thought provoking and
likely to engage many students in out-of-class discussions
(e.g., “What are the consequences of performance-enhancing
drugs?,” which is coupled with another ELSI: “If pills could
make you remember or forget, would you take them?”). They
are similar in construction to those seen in typical in-
троductory texts, but they are far more central to the design
of the course. Prominence and use generates student discus-
sion; we believe that is why the ELSIs are likely to succeed
and are an unflinching part of the approach taken in ICB—
they are not an afterthought or simple enrichment exercises.

The text is written in a more colloquial style than is typical.
The authors state that they hope this will improve accessi-
bility and student engagement in the reading. We differed
in our views as to whether this is appropriate. Students will
likely find the informal, less jargon-burdened text easy to

Figure 1. Graphics allow students to discern the relationship between material presently under consideration (green) with regard to the five “big ideas” (yellow) and level of the biological hierarchy (blue). Both examples are for chapters that center on information. The organization fits a two-semester course focusing first on molecular, cellular, and organismal biology (top three rows) and then on organisms and ecology (lower three rows), with each semester beginning with information and progressing to emergent properties. (A) The graphic for chapter 1, “Information in Molecules and Cells.” (B) The graphic for chapter 3, “Reproduction and Cell Division.”
read. But one can argue that learning to read more difficult texts is a skill that should begin early in a student’s career. And, although reduction in the number of terms introductory students encounter seems to be a worthy goal, what is appropriate for one instructional setting might be less so for another.

The use of hypertext and illustration is very effective. The illustrations are clear. They serve to advance points made by the text and are helpful in developing a student’s ability to visualize processes (e.g., what happens when a three-dimensional object is shown as sliced sections). Hypertext was used to give definitions of all important terms (click on the word and see the definition); to lead a student to connected sections for review or preview; and to effectively link the text with Web-based resources such as simulations, animations, movies, and appropriate examples of primary literature (although some of these articles were not open access). The number and types of hyperlinks represented a good balance between assisting informed student discovery without simply revealing the answers. One small complaint is that it would be useful if the elements of the navigation graphic (Figure 1) were electronically linked to the corresponding chapters. Additional comments on the use of hypertext and general functionality of the e-text can be found in the Supplemental Material.

The presentation of material was, for the most part, clear and accurate. Some (perhaps inevitable) issues of accuracy or emphasis can nevertheless be found. For example, in chapter 9 (“Neurons and Muscles”) the text seems (incorrectly) to attribute resting potentials to the electrogenic nature of the Na+/K+-ATPase—whereas a linked animation correctly explains the central role of membrane permeability and ion concentration differences. Another concern arises from the designation of the term “homeostasis” as a unifying theme. Homeostasis requires biological mechanisms that serve to confine the states of some variable to a small range of values in the face of a wider range of environmental perturbation. Chapter 29 (“Population Homeostasis”) seems to lead students to conclude that population sizes are regulated homeostatically through density-dependent effects on births and deaths. Most population ecologists endorse the notion of density-dependent “equilibrium” population sizes (usually called carrying capacities) that exist for a specific set of environmental conditions. But change the broad environmental conditions and the carrying capacity immediately changes (think of the history of human populations and technologies), unlike what is observed in homeostasis. We suggest that a more globally unifying “big idea” is regulation; this broader term better allows concepts such as equilibrium, regulation, and homeostasis (a relatively uncommon condition) to be explored and distinguished from one another.

In a few other places, we disagree with the authors’ approach to a topic. For example, in the treatment of cellular energetics, we find that it is more useful to focus on Gibbs free energy and its relationship to displacement from equilibrium than on bond energies of molecules. The free-energy approach is especially useful, because bioenergetics subsumes a broad array of processes beyond chemical reactions such as ion and electrical gradients and mechanical processes. Moreover, use of this approach facilitates (and requires) that students gain a conceptual facility with equilibria (of all types) and steady states. We recommend a short essay by Richard Feynman (2008) as an accessible starting point to a guided discovery of these important topics.

As with any endeavor, ICB is a product of the vision of its authors, inevitably influenced by their training and expertise. The great strengths of ICB are at the cellular, molecular, and ecological levels of the biological hierarchy, reflecting the authors’ biological foci and the reality that most biological research is presently focused on the “small” and the “big.” The authors made an honest attempt to minimize the small-versus-large divide by seeking common themes and providing for exploration of system- and organism-level mechanistic biology (see Figure 1). But it is clear that their experience (and most successful presentation) is elsewhere. For example, chapter 9 (“Neurons and Muscles”), which was intended to emphasize organisms, does an excellent job in guiding students through research on the regulation of synapse formation at the cellular and genomic levels. But the highest level of structural complexity discussed is between one cell and another. The authors miss a chance to explore emergent properties found in networks of connected neurons. The complexity and (often) flexibility of these networks allow neural computations related to perception (e.g., vision) and much of cognition and learning. Likewise, treatments of plant and animal physiology (chapters 12 and 15) focus heavily on processes at the cellular or molecular level. Chapter 28 (organism homeostasis) presents some interesting data regarding mammalian thermoregulation but without mentioning the roles of metabolism and circulation in temperature regulation. Temperature regulation is a subject that is often used to illustrate how interacting organ systems can achieve varying degrees of regulation. It can be used to briefly explore the salient features of these systems (“What must a pump do?” and “How is flow controlled?”) and is an ideal subject for mathematical treatments (applications of Newton’s law of cooling, surface, volume, and size) and seeing how principles from physics (e.g., diffusion and bulk flow of heat) apply to organisms. A balanced grounding in tissue/system/organism function is important to the education of beginning students.

All textbooks have problems such as the ones just mentioned. We are concerned that no content reviewers are acknowledged in the prefatory material of this text, and Trinity uses the process of “assisted publication,” so it may be that the text was not subjected to as thorough a review as a traditional textbook backed by the resources of a major publishing house. However, since the electronic format permits updating at the authors’ pleasure, mistakes can be easily corrected as users offer feedback.

ICB, OTHER INTRODUCTORY TEXTS, AND THE PARABLE OF THE BLIND MEN AND THE ELEPHANT

Each biologist has his or her own perspective as to how students should be introduced to biology, much as in the ancient parable wherein each wise but blind man attributed different characteristics to the elephant. One approach is to admit that other views are useful but, nevertheless, select a particular perspective. The three BSCS high school biology texts launched in the 1960s are good early examples of this
The more common path has been to examine the "elephant" from every possible perspective and pedagogy and produce a textbook that theoretically would meet all needs. Given the costs of traditional publication, there is logic in producing a book that has the greatest adoption potential. Thus, traditional textbook writers have compiled increasingly massive compendia, with multiple authors and expert advisors, richly detailed illustrations, numerous study questions, boxed features on applications and ethical issues, and examples of applications of mathematics and chemistry, along with an array of electronic resources and instructional aids. For example, the widely adopted introductory biology textbook that we use at Holy Cross (Sadava et al., 2014) has more than 1300 pages and uses multiple authors and advisors, each an expert on some range of topics.

This thoroughness is often overwhelming for both students and faculty members. Students become lost in the detail and bombardment of terms and fall prey to frustration or boredom. Instructors provide study guides to help students focus, leading students to minimize their reading and concentrate on slides presented in class, which reflect what the instructors believe to be most important.

ICB was partially designed as an antidote to the expansive text, and it effectively facilitates a path to teaching and learning that takes advantage of recent educational research promoting the benefits of student-centered active learning. For these reasons and because of its inherent strengths as a text, it is a significant contribution. But is it for everyone? The answer, of course, is no. We have a different take on what topics should be presented and how. For instance, we believe that it is not just theory and experiment that motivate students; many thrive and grow from knowledge of biological diversity. Other differences that matter to us, such as how to explain biological energetics or the importance of good treatments of organismal biology, we touched upon earlier. We also place more importance in the value of terminology than do the authors of ICB. So, at our institution, we are moving along a largely parallel but different path as we strive to make education more exploratory, active, and integrative—a path that owes much to that advocated by Freeman et al. (2011). We look for textbooks with treatments that include these perspectives. A second issue is how the introductory course fits into the entire department's curriculum. Coadaptation of introductory and advanced courses is required for a coherent student experience.

These quibbles and reservations aside, we believe that in the hands of motivated and skilled instructors, ICB represents the core of an approach that is likely to excite and engage students more than does a more traditional course. The text and supplemental materials are effective tools to anchor courses of guided self-discovery of biology and the scientific process. This approach is more likely to motivate students to develop a usable knowledge foundation and problem-solving skills, and it should also facilitate the development of their metacognitive skills (NRC, 2000). A study done at Davidson College the first year that the ICB version of first-semester introductory biology course was offered in parallel with a more traditional version showed that student performance was similar across approaches on factual content but that ICB students performed better and showed more improvement on tasks involving data interpretation and had a better conceptual understanding of biology as a discipline (Barsoum et al., 2013). Moreover, even educators who choose not to adopt ICB would likely benefit from studying its intent and execution. It exemplifies an innovative and enlightened new direction in education that is worthy of emulation by those who see the biology elephant with different perspectives. If we may conclude with a personal reaction, reviewer K.N.P. would have been overjoyed to have experienced the ICB approach when, a mere four-plus decades ago, he took introductory biology in the very same rooms as the students of Campbell, Heyer, and Paradise do today!

ACKNOWLEDGMENTS

We thank Mary Lee Ledbetter for many useful stylistic suggestions.

REFERENCES

American Association for the Advancement of Science (2011). Vision and Change in Undergraduate Biology Education: A Call to Action, Washington, DC.


NEST 2014: Views from the Trainees—Talking About What Matters in Efforts to Diversify the STEM Workforce

Andrew G. Campbell, * Rachel Skvirsky, † Henry Wortis, ‡ Sheila Thomas, § Ichiro Kawachi, / and Christine Hohmann ¶

*Department of Molecular Microbiology & Immunology, Brown IMSD Program, and Brown MARC Program, Division of Biology & Medicine, Brown University, Providence, RI 02912; † IMSD Program, College of Science & Mathematics, University of Massachusetts Boston, Boston, MA 02125-3393; ‡ Department of Integrative Physiology & Pathobiology, Graduate Program in Immunology, and Tufts PREP Program, Tufts University School of Medicine and the Sackler School of Graduate Biomedical Sciences, Boston, MA 02111; § Harvard Graduate School of Arts & Sciences Holyoke Center, Cambridge, MA 02138; / Department of Social and Behavioral Sciences, Harvard IMSD Program, Harvard School of Public Health, Boston, MA 02115; ¶ Department of Biology and RISE Program, Morgan State University, Baltimore, MD 21251

BACKGROUND AND INTRODUCTION

Efforts to diversify the U.S. science, technology, engineering, and mathematics (STEM) workforce have been led by various stakeholders across all disciplines but most notably by the funding agencies and by the trainers (National Research Council [NRC], 2007, 2011; Tabak and Collins, 2011; Wilder et al., 2013). However, missing from this work and these conversations are the voices of the trainees at all levels. While this work has now begun to include the views and opinions of the postdoctoral community (www.nationalpostdoc.org), rarely does it involve trainees at mid- and entry levels of the pipeline. Interest in faculty careers decreases as training progresses (Gibbs and Griffin, 2013). Additionally, given that the greatest diversity in the scientific community is found at the undergraduate level, followed by the postbaccalaureate and graduate levels (Ramdial and Campbell, 2014), there is an urgent need to capture these unfiltered viewpoints that form the foundations for career decisions and actions.

STEM training program analyses aimed at defining what matters in trainee choices, persistence, and motivation have always been guided from the top. Part of this work relies on administering surveys constructed using assumptions and inferences that we as trainers make regarding what motivates trainees and the factors that affect their choices. While useful, these approaches are often derived in prescriptive ways, which can lead to unintended biases by undervaluing or failing to measure the traits and the attributes trainees themselves possess and value, attributes that could be beneficial in contemporary interdisciplinary science. The issue of career choice, persistence, and motivation is a complex matter; discussion should not be limited by top-down decision making or by overly structured theoretical frameworks. Trainee choices are shaped by internal decisions as well as by external factors (Skaalvik and Skaalvik, 2002) that are not always apparent or understood. These may go unrecognized, because adequate time is not given to trainees for reflection. With regard to knowing what matters, trainees must have adequate opportunities to think deeply and reflect on what is important to them and what motivates them most in pursuing STEM careers.

In spite of programmatic investments made over the past 40 yr, only modest gains have been seen in the number of underrepresented minority members (URMs) who join the STEM workforce (Mervis, 2006). Given that program features and training practices designed to increase URM entry into the STEM pipeline and workforce mirror those provided to their non-URM peers, the modest outcome...
achieved to date is disappointing. This outcome indicates that other elements essential for URM trainee recruitment and success may be missing from current training programs, which in turn suggests the need to empower trainees with the agency to contribute to the redesign of the STEM training pipeline by providing them with the opportunity for greater input into discussions aimed at improving practice to achieve better outcomes.

The Northeast Scientific Training (NEST) Programs Retreat was established with three purposes in mind: 1) to provide URM trainees with a “community of scholars retreat setting” as an alternative to the formal and highly structured experience of scientific conferences; 2) to create a venue and environment free from daily distractions, one in which trainees could meet peers and near peers along the entire training spectrum to informally discuss, inform, and question one another about careers, career paths, and choices; and 3) to give trainees the opportunity to formally report their concerns, desires, and recommendations on how to increase URM student involvement in STEM fields and careers. These purposes are linked to the goals of 1) broadening the trainees’ views and understanding of the lives of scientists and scholars, and 2) providing trainees with a sense of purpose and empowerment as stakeholders in the future of science and of education in the United States. Figure 1 presents an outline of the 2-d NEST Retreat.

**ORGANIZATION OF DISCUSSION GROUPS**

Participants were assigned to working groups that were fixed for the duration of the retreat; each group included a mix of undergraduates, graduates, postbaccalaureate students, and postdoctoral fellows, as well as one to two faculty/staff. Initial group meetings took place in the absence of faculty/staff, allowing trainees to get to know one another and engage in peer and near-peer mentoring. These trainee-only sessions, moderated by peer leaders selected by peers from within each group, helped to maximize interactions and discussions about shared experiences and knowledge. Furthermore, the trainee-only group sessions helped 1) trainees speak freely with one another on topics and concerns they would not raise in the presence of faculty, 2) develop trainee consensus around ideas and opinions regarding training by drawing on contemporary and relevant examples from their own experiences, 3) minimize organizer control of the pace and direction of discussions, and 4) establish the parameters of the subsequent full group’s work over the course of the retreat. The trainee-only group sessions also allowed trainees to preview the open-ended topics and questions, which were the centerpieces for full-group (groups involving faculty/staff) discussions that would take place over the day and a half that followed.

---

**Figure 1.** NEST Retreat Program of Activities Daily activities of the day and a half long retreat are presented. Attendees arrived late Friday afternoon of day 1 and departed at noon on day 3. Participants were faculty members, most of whom were training program directors, and trainees, including undergraduate students, graduate students, and postdoctoral fellows. “Attendees” refers to retreat participants and members of the organizing committee.
The morning of the first full day of the retreat was devoted to full-group discussions facilitated by faculty and staff. Faculty and staff were excluded from groups in which trainees from their own institutions were present. During group sessions, participants discussed the open-ended topic “Understanding, Deciphering and Re-imagining the Pathways to Training and Scientific Careers.” For the afternoon sessions, the same groups were reconvened, so participants could respond to the following two questions: 1) What are the traits of a good trainee and a good trainer? 2) How do you measure trainee success? For each session, trainees and facilitators summarized in writing the outcomes of their discussions. At the end of the day, one peer-elected trainee leader representing each group was selected to report to all retreat attendees the outcomes and actionable points expressed by their group.

MEETING ATTENDEES

Admission to the 2014 NEST Retreat was competitive and open to all; preference was granted to trainees involved in structured, federally funded, research training programs. Faculty and staff were similarly selected, with preference given to those demonstratively engaged in STEM student training. Attendees came from diverse backgrounds and academic institutions spanning the Carnegie classification system. These included majority-serving institutions, historically black colleges and universities, and Hispanic-serving institutions, large and small, as well as private and public.

To ensure that pertinent student viewpoints and experiences were captured, the organizers sought attendees who represented programs at their home institutions that function specifically to increase the participation of individuals from underrepresented groups in the STEM disciplines. Faculty and staff in attendance were also drawn from these programs. Sixteen trainers and 50 trainees from colleges and universities representing undergraduate, graduate, postbaccalaureate, and postdoctoral training programs were invited to participate in a 1½-d retreat at the Marine Biological Laboratory in Woods Hole, MA. More than half of the trainees were students at various stages of undergraduate training. Most attendees were from the Northeast, representing Brown University, Harvard University, Morgan State University, Northeastern University, Tufts University, the University of Massachusetts Boston, Cornell University, the University of Buffalo, LaGuardia Community College, Pine Manor College, and the College of Mount St. Vincent. Additional attendees represented the University of Michigan, Arizona State University, the Universidad Metropolitana, and Elizabeth City State University. Trainees and trainers present represented the life, biomedical, public health, and engineering sciences. Nontrainees in attendance included faculty and nonfaculty program directors and high-level university administrators.

OBSERVATIONS AND OUTCOMES

The first observation noted at the start of the retreat was that trainees expressed satisfaction in being able to dress casually, which contrasted with their perceived need for a formal dress code at scientific meetings. Some expressed that the ability to dress casually was less constraining and meant they could bring their authentic selves to the retreat.

The retreat began with participants engaging in a competitive exercise in which they were asked to meet as many of the other participants as possible, learning their names and one unique STEM-related feature of the individual, and recording that information. Those who most successfully completed the exercise, which took place informally over the day-and-a-half retreat, received a prize.

Trainees attending the retreat were invited to bring a poster of their most recent research work to display for the duration of the retreat. Trainees presented the details of their work in these posters during two early-evening poster sessions. The poster sessions, which followed late afternoon socials, were held on the first and second evenings of the retreat. These sessions were useful, as they created informal galvanizing conversational settings primarily for trainees and displayed the broad range of scientific work in which trainees are involved. The majority of the retreat’s work was accomplished in small groups, which met initially the first evening of the retreat for 2 h. These groups met during the following day for two 3-h group sessions followed by a 1-h evening plenary session in which representatives shared the outcomes of discussions. The open-ended nature of the retreat’s discussion topic and questions fostered responses that were directly or indirectly related to the topics and questions posed. Reporting out provided trainees with opportunities to share with their peers and faculty/staff the outcomes of their discussions and to highlight the issues that resonated with them the most. This activity also allowed the process and considerations of career interest formation, desires, and concerns of each group to be shared with all groups.

The final morning of the retreat was dedicated to an overall summary of the retreat discussions that had taken place over the past day and a half, including the previous all-attendee session. This final morning session allowed for clarifications to be volunteered, corrections made, final questions asked, and consensus to be reached about what matters most to trainees, impacting their entry, persistence, and motivation to pursue STEM careers. Here we summarize the outcomes of this reporting out, framed as actionable items:

1. Adapt STEM training to include or make room for a social justice component. Trainees expressed the desire for opportunities to do science with a purposeful social justice component, a desire that does not preclude performing traditional bona fide research at the highest level. This desire appears to reflect the sense of disconnect and marginalization that trainees feel within the academy and the scientific community. It also appears to align with their concerns for issues such as health disparities, which are evident for underrepresented/disadvantaged groups. Coincidentally, this issue was recently a featured outreach topic in ASMB Today, which is published by the American Society for Biochemistry and Molecular Biology (Thompson et al., 2014), indicating that, like so many other issues raised at the retreat, interest in this issue is not unique to URM students.

2. Assist us in our desire to better communicate science to nonscientists. Trainees expressed a desire to have their STEM training experience prepare them to communicate science more broadly to nonscientists. This desire appears
to stem from the fact that the families of most trainees have a good understanding of (for example) what a physician does, but very little understanding of what a scientist does. This lack of awareness likely reflects inadequate communication from the science community to the larger world. By increasing public communication and education outreach, STEM trainees, especially those from underrepresented groups, may be able to build greater social and intellectual capital with their communities and thus garner the greater family support many need to persist and succeed in science.

As is the case with inclusion of a social justice perspective, the desire for better communication skills does not preclude performing bona fide research at the highest level. It again suggests the sense of isolation that trainees feel from the broader “lay” community. The failure of scientists to reach these and other underserved communities may create an unintended perception of elitism.

3. STEM to STEAM and beyond. Support trainee desires for interdisciplinary cross-talk and training. Trainees recognize the growing need for communication between the physical, life, and biomedical sciences and feel that this is important for the advancement of science. However, they feel that their interest in broader and truly interdisciplinary cross-talk and communication is neither supported or recognized and that little room exists for this type of interest. For them, there is a strong desire for STEM and non-STEM communication and cross-talk. The concept of STEM trainees and practitioners embracing non-STEM fields is a concept promoted by many, including John Maeda, president of the Rhode Island School of Design (http://stemtosteam.org). The premise of this concept is to have STEM trainees value the potential contributions of art and design to the sciences. STEM plus the arts transforms STEM to STEAM, which embodies the vision of interdisciplinary cross talk, training, and practice.

4. Educate us earlier rather than at the late graduate and postdoctoral levels about science careers. Trainees felt that their current training experiences should provide more information earlier about the world of research and research career options and paths that can help them understand better, and sooner, all of the things that they can do with a PhD. This desire is not a novel one. What is different, however, is that the current efforts for career planning appear to occur late in the training process, while trainees felt that they needed career information earlier in their training to make the most informed career choices.

5. Give better guidance and assistance in achieving work–family life balance. Trainees expressed that they needed to be informed earlier about how to balance family life with the demanding life of a research scientist. This response affirms that mid- and entry-level trainees in the pipeline grapple with the same issues more senior trainees face. Current efforts in this area, through programs like ADVANCE (National Science Foundation) and other similar programs, are focused on postdoctoral fellows transitioning to faculty positions. More junior trainees, including those at the undergraduate level, face similar and immediate challenges around balancing family life and training life issues. Their challenges may not all be related to childbearing and child rearing, but may include responsibilities for raising siblings and supporting their families in a nonparental but equally essential family role that oftentimes includes managing financial challenges.

6. Re-evaluate the current metrics that fail to value diverse traits trainees can bring to science that may benefit science. Trainees felt that some of their strong attributes are not being utilized or valued in evaluating them as prospective trainees and scientists. These included their capacity to think in truly interdisciplinary ways—connecting STEM field disciplines to other disciplines, including their creativity (all elements of STEAM); and their responsibility, organizational skills, ability to be good listeners, and capacity to develop independence. They felt that these traits, which added to the wholeness of being a scientist, continue to be undervalued. At the same time, trainees expressed concern that gender and other biases persist and that these continue as the by-products of insufficient interaction between trainers and trainees from diverse backgrounds. Cultural differences are still misunderstood in ways that may lead trainees to be perceived as lazy or as not enjoying science. The short summer training experiences these trainees have in labs at research-intensive institutions are not sufficient to address this problem.

7. Provide access to invested mentors and graduate school guidance. Trainees continue to feel that they do not have access to invested mentors who show a genuine interest in their careers. They felt that more work is needed to align trainee and trainer expectations. Trainees also expressed that they did not know what their goals should be for a given career path nor did they understand their mentors’ goals.

8. Create opportunities for ancillary training. Some trainees indicated that greater definition of their areas of STEM career interests came about only after a series of meaningful experiences, at which point they might be 2 yr into an undergraduate or PhD program. The training process, however, committed them to paths that might not completely align with the preferred career interests they developed. These trainees felt that the only way to access those areas that truly interested them was by exiting their current training pathways. Others, who did not recognize some of the options available to them, would simply choose to exit the STEM fields altogether. Trainees asked for solutions to help better align their interests with the training process. Programs are clearly needed to meet this important need. One solution may be the pilot program launched by Brown University (2014), the Open Graduate Curriculum (www.brown.edu/academics/gradschool/opengraduateducation). This program allows current PhD trainees in one discipline to enroll in a second graduate program to earn a master’s degree, thereby receiving training in two complementary areas. A PhD student in pathobiology, for example, whose research focus in cancer biology research may be very basic in nature, may come to realize that his/her interest in cancer research overlaps epidemiology and public health. As part of the Open Graduate Curriculum, such a student would remain in his or her PhD program but could now enroll in the Master’s of Public Health Program, resulting in training synergy. Supporting access to similar dual-degree programs accommodates these interests in early ancillary training. Accommodating these interests also has
the potential to lead to the birth and development of new interdisciplinary and emerging fields of study.

RELATIONSHIP OF RECOMMENDATIONS TO THE CURRENT TRAINING FRAMEWORK

Efforts to improve STEM diversity over the past 40 yr have been centered on identifying the key correlates for trainee success at the undergraduate, graduate, postdoctoral, and faculty levels and on providing support accordingly (NRC, 2005, 2011). Numerous published studies seek to define the contexts that promote diversity and engagement leading to pursuit of a science career by underrepresented individuals (Hurtado et al., 2009; McGee et al., 2012). These and other works have contributed to themes that have become universally recognized as critical for trainee success. Critical themes include helping trainees to understand the culture of science; to manage racial and social stigmas; and to develop scientific identity, self-efficacy, and motivation. The persistent deficit of URMs in STEM careers also validates the continued need for structured programs that create opportunities for student engagement in research activities. However, given the poor progress in diversifying the pipeline population and workforce thus far, it is clear that the current architecture and format of these programs is not sufficient for addressing the problem. More nuanced but equally important early-career trainee needs appear to have gone unaddressed, including some of those framed by the recommendations made in this report. It should, of course, be noted that the current report is based on a dialogue with a small sample of URM students (∼50) and these ideas need to be vetted with a larger pool of students. When combined with the current knowledge of and efforts to implement intervention practices, the menu of actionable items presented here may provide the critical elements necessary to bridge programmatic and individual STEM aspirations. These are presented in Figure 2 as part of the revised training timeline for STEM trainees. Adopting these recommendations will likely be beneficial, as they would likely strengthen the current frameworks of the various intervention models and program practices designed to improve STEM field diversity, trainee persistence, and success.

LESSONS LEARNED

The NEST Retreat was designed to support a “discovery-driven” approach to training and program development. Many of the issues raised at NEST 2014 are familiar issues, such as those related to family life–work life balance, the potential to lead to the birth and development of new interdisciplinary and emerging fields of study.

Figure 2. The STEM training timeline and accompanying supporting activities. The training timeline moves from left to right; an arrowhead represents the end of one level of training. Standard curricular training at the undergraduate and graduate levels entails completion of “for-credit course work,” indicated by the solid black arrows. Research training at the undergraduate, graduate, and postdoctoral levels is shown by the open/transparent arrows. The dashed-line, shaded arrow represents additional and often optional curricular training. Postbaccalaureate training, which occurs between the undergraduate and graduate training periods, is not shown but resembles training received in the terminal undergraduate year. An approximation of the start and duration of the current typical set of supporting activities that accompany formal degree and postdoctoral training is shown above the training line. The proposed revised and reimagined timelines for these supporting activities are shown below the training levels.
mentoring, and career pathways. What is different, however, is the early training levels at which they have been reported to be important, indicating the impact of these issues on entry-level trainees in making decisions about entering the pipeline and persisting to advanced levels. Many of the recommendations made can be addressed by integrating supporting practices into the standard plans at the undergraduate, graduate, and postdoctoral levels, without altering curricular or research expectations of trainees. For example, some of these supporting activities can be accomplished by providing trainees with non–credit bearing short-term educational training modules that deliver the necessary content to respond to trainee needs, while supporting their STEM persistence and success. Modules could be offered sequentially at times that do not conflict with the current training or educational programs. Each module could provide intensive training sessions with 10–12 contact hours offered over a 1- to 2-wk period and could be cotaught by faculty and near peers who have had recent experiences in the area (Thompson and Campbell, 2013).

The concepts of including social justice considerations and communication training have never been central to the purview of STEM field scientists. These interests among current trainees may be a by-product of how the application of science in society differs today from its application in past decades and who this application impacts. In addition to sharing their views at the retreat, one group of trainees submitted written comments and summaries of their retreat experiences (Supplemental Material).

The outcomes of this first retreat yielded insights we believe would have gone undetected at traditional scientific gatherings. We expect that much of what has been learned will be incorporated into the training programs that many of us lead as program directors, faculty, and staff. In summary, the overall assessment of NEST 2014 is that much of what was learned points to the dynamic nature of the STEM training pipeline, which we have historically treated as a relatively static structure. Trainee feedback and input can and should stimulate pipeline change. Responding to many of the issues raised above can help us to reimagine the training pipeline as a structure that bends to better address and adapt to trainee interests and that broadens to help trainees to broaden their skill base and to create new training modalities that better serve the STEM disciplines.

ACKNOWLEDGMENTS

The authors thank Ms. Jennifer Daly for administrative and editorial assistance and Ms. Karen Ball for administrative assistance. We also thank Liza Cariaga-Lo, Laura Liscum, and Elizabeth Solomon for their roles as group facilitators. The authors are indebted to all of the trainees who attended the 2014 NEST Retreat and made this report possible by sharing their views. The report was assembled by and represents the interpretation of the authors. This article and the NEST Retreat were supported by the National Institute of General Medical Sciences of the National Institutes of Health under award numbers R13GM106577 and R25GM083270 to A.G.C., 5R25GM058904 to C.H., R25GM076321 to R.S., 2R25GM055353-13 to I.K., and T32AI007077 and R25 GM066567 to H.W.

REFERENCES


Gibbs KD, Griffin KA (2013). What do I want to be with my PhD? The roles of personal values and structural dynamics in shaping the career interests of recent biomedical science PhD graduates. CBE Life Sci Educ 12, 711–723.


Essay

Research-Based Implementation of Peer Instruction: A Literature Review

Trisha Vickrey, Kaitlyn Rosploch, Reihaneh Rahmanian, Matthew Pilarz, and Marilyne Stains

Department of Chemistry, University of Nebraska–Lincoln, Lincoln, NE 68588

Submitted November 6, 2014; Revised December 7, 2014; Accepted December 10, 2014
Monitoring Editor: Daron Barnard

Current instructional reforms in undergraduate science, technology, engineering, and mathematics (STEM) courses have focused on enhancing adoption of evidence-based instructional practices among STEM faculty members. These practices have been empirically demonstrated to enhance student learning and attitudes. However, research indicates that instructors often adapt rather than adopt practices, unknowingly compromising their effectiveness. Thus, there is a need to raise awareness of the research-based implementation of these practices, develop fidelity of implementation protocols to understand adaptations being made, and ultimately characterize the true impact of reform efforts based on these practices. Peer instruction (PI) is an example of an evidence-based instructional practice that consists of asking students conceptual questions during class time and collecting their answers via clickers or response cards. Extensive research has been conducted by physics and biology education researchers to evaluate the effectiveness of this practice and to better understand the intricacies of its implementation. PI has also been investigated in other disciplines, such as chemistry and computer science. This article reviews and summarizes these various bodies of research and provides instructors and researchers with a research-based model for the effective implementation of PI. Limitations of current studies and recommendations for future empirical inquiries are also provided.

INTRODUCTION AND BACKGROUND

Discipline-based education researchers have responded to calls (President’s Council of Advisors on Science and Technology [PCAST], 2010, 2012) for instructional reforms at the postsecondary level by developing and testing new instructional pedagogies grounded in research on the science of learning (Handelsman et al., 2004; National Research Council [NRC], 2011, 2012). These research-based pedagogies significantly increase both student learning and attitudes toward science (NRC, 2011, 2012). Peer instruction (PI), which was first introduced by Eric Mazur in 1991 (Mazur, 1997), is an example of a research-based pedagogy. In PI, traditional lecture is intermixed with conceptual questions targeting student misconceptions. Following a mini-lecture, students are asked to answer a conceptual question individually and vote using either a flash card or a personal response system commonly called a “clicker.” If a majority of students respond incorrectly, the instructor then asks students to convince their neighbors that they have the right answer. Following peer discussion, students are asked to vote again. Finally, the instructor explains the correct and incorrect answers (Mazur, 1997; Crouch and Mazur, 2001). It is important to note that, although PI is commonly associated with clickers and there have been helpful reviews on best practices for clicker use (Caldwell, 2007; MacArthur et al., 2011), this article is focused on PI, a specific, evidence-based pedagogy that can be effectively implemented with or without clickers.

PI has been primarily disseminated and adopted by physics instructors (Henderson et al., 2012) but has also been
widely adopted by faculty members in the biological sciences and other science, technology, engineering, and math (STEM) fields (Borrego et al., 2011). However, a recent study indicates that instructors adapt the PI model when implementing it in their classrooms, often eliminating either the individual voting or peer discussion steps, which are critical to the effectiveness of the pedagogy (Turpen and Finkelstein, 2009). Modification of evidence-based instructional practices has been associated with reduced learning gains in other studies (Andrews et al., 2011; Henderson et al., 2012; Chase et al., 2013). While sometimes necessary, modifications are often made without fully realizing how they will impact effectiveness. The lack of knowledge of adequate ways to adapt a practice is due to two major reasons: 1) for most evidence-based instructional practices, few empirical studies have been conducted to identify the critical elements of the practice that make them effective; 2) for the few practices for which this type of research exists, the studies have been reported in various fields and journals, making it difficult for instructors and researchers alike to have a comprehensive, research-based description of the most effective implementation. PI falls into the latter group. There have been numerous studies exploring various aspects of the PI model, but these studies have been disseminated in a variety of ways. This article is intended to provide a comprehensive review of these studies. In particular, we summarize studies demonstrating the effectiveness of PI, describing the stakeholders’ views on PI, and identifying the critical aspects of PI implementation. We foresee that this article will be used by researchers to design instruments that measure fidelity of implementation of PI, professionals involved in professional development to provide them with resources for their sessions, and instructors from multiple STEM disciplines interested in implementing this practice.

METHOD
The search of articles for this literature review was constrained by the following parameters: 1) studies had to be conducted at the college level and in STEM courses; 2) studies had to report results that could only be attributed to the implementation of PI; 3) studies in which PI was implemented as part of a set of several evidence-based instructional practices and that only provided results for the set of practices were not included; and 4) implementation of PI had to follow the steps described by Mazur (1997), which have been associated with measurable learning gains (see Impact of PI on Students). Each step is discussed in more detail in the Evidence-Based Implementation of PI section. PI was used in combination with the following keywords in the ERIC, Web of Science, and Google Scholars databases: learning gains, retention, flash cards, clickers, personal response system, problem solving, concept inventory, concept test, voting, histogram, and peer discussion. The studies that met the criteria for this literature review are included in Supplemental Table 1.

IMPACT OF PI ON STUDENTS
Studies have measured the impact of PI on learning gains, problem-solving skills, and student retention.

Are There Measurable Learning Gains with the Use of PI?
The impact of PI on student learning has been most commonly measured in physics through the calculation of normalized learning gains. Normalized learning gains were first introduced by Hake (1998) in a widely cited study demonstrating the positive impact of active-learning instruction in comparison with traditional lecture. Normalized learning gains are calculated when a conceptual test, typically a concept inventory (Richardson, 2005), is implemented both at the beginning and end of a semester/unit/chapter. The actual gain in a student’s score is divided by the maximal possible gain, \( \frac{(\text{posttest} - \text{pretest})}{(100 - \text{pretest}) \times 100} \), which allows a valid comparison of gains between students with different pretest scores. In a longitudinal study, Crouch and Mazur (2001) explored the impact of PI compared with traditional lecture on student learning in algebra- and calculus-based introductory physics courses at Harvard University. At the beginning and end of a semester, they administered a conceptual test, the Force Concept Inventory (FCI; Hestenes et al., 1992), to measure changes in normalized learning gains as they implemented either the PI pedagogy alone or a combination of PI and just-in-time teaching (Novak, 1999; Simkins and Maler, 2009) pedagogies. During the 10 yr of data collection, Crouch and Mazur (2001) observed normalized learning gains that were regularly twice as large as those observed with traditional lecture, even when implementing PI alone.

To further validate the positive impact of PI on student learning, these authors collected survey data from other current and past implementers of PI who had administered the FCI (Fagen et al., 2002). The survey was posted on the Project Galileo website and directly emailed to more than 2700 instructors. The authors identified 384 instructors who were current or former PI users. They were able to obtain matched pre–post FCI data from 108 of these instructors representing 11 different institutions, including 2-yr, 4-yr, and research-intensive institutions, and 30 different courses. In 90% of these courses, they found medium normalized learning gains (medium \( g \) ranges from 0.30 to 0.70) with only three courses falling below that range. According to Hake’s (1998) study, medium normalized learning gains are typically not achieved in traditionally taught courses. Another study by Lasry et al. (2008) compared the impact of the first implementation of PI in physics courses at Harvard University with that of implementation at a 2-yr college at which student’s preinstructional background in physics is lower. Their quasi-experimental study demonstrated that students in these two settings achieved similar normalized learning gains (\( g = 0.50 \) at Harvard University and \( g = 0.49 \) at the 2-yr college).

The impact of PI on learning has been studied in disciplines other than physics as it has gained popularity. In the geosciences, McConnell et al. (2006) determined that the average difference between post- and pretest scores on the Geosciences Concepts Inventory (GCI; Libarkin and Anderson, 2005) was greater with PI pedagogy, and Mora (2010) reported greater normalized learning gains on the GCI compared with traditional lecture. Moreover, students in an introductory computer science course implementing PI scored half a letter grade higher on a final examination compared with peers in a lecture-based course covering the same topics.
Does PI Improve Problem-Solving Skills?

Studies have also focused on characterizing the impact of PI on students’ problem-solving skills. In a study by Cortright et al. (2005), PI was introduced in an exercise physiology course. Students were randomly assigned into a PI group or a non-PI group in which students were presented with the in-class concept test but were instructed to answer the questions individually rather than discussing them with their peers. Students in the PI group improved significantly ($p = 0.02$) in their ability to answer questions designed to measure mastery of the material. Importantly, the PI group’s ability to solve novel problems (i.e., transfer knowledge) was significantly greater compared with that of the non-PI group (Cortright et al., 2005).

In another study, PI was introduced in a veterinary physiology course (Giuliodori et al., 2006). Giuliodori and colleagues compared student responses before and after peer discussion to determine whether or not PI improved students’ ability to solve problems requiring qualitative predictions (increase/decrease/no change) about perturbations to physiological response systems (i.e., integration of multiple concepts and transfer ability). The number of students correctly answering questions improved significantly after peer discussion (Giuliodori et al., 2006). Moreover, in a comparison of ability to transfer knowledge, students in an entomology course for nonmajors using PI scored significantly higher ($p < 0.05$) on a near-transfer task (e.g., application of prior learning to a slightly different situation) compared with students in a non-PI group (Jones et al., 2012). These studies suggest that PI improves students’ ability to apply material to novel problems.

In addition to improvements on multiple-choice and qualitative questions, PI has been associated with learning gains on quantitative questions (Crouch and Mazur, 2001). The previously mentioned longitudinal study conducted at Harvard University in physics courses compared quantitative problem solving in courses with and without PI. A final examination consisting entirely of quantitative problems was administered after the first year of instruction with PI. The mean score on the exam was statistically significantly higher in the course with PI compared with traditional lecture. Thus, PI pedagogy can enhance both qualitative and quantitative problem-solving skills.

How Does PI Affect Attrition Rates?

Students’ persistence in STEM fields is a critical concern at the forefront of federal and national initiatives (National Science Foundation, 2010; PCAST, 2012). Several studies have examined student retention rate in courses using PI. In the instructor survey study conducted by Mazur and colleagues, instructors implementing PI reported lower student attrition rates compared with those using traditional lecture (Crouch and Watkins, 2007). In the study comparing the implementation of PI in physics courses at Harvard University and at a 2-yr college (Lasry et al., 2008), the dropout rate (difference between the number of students enrolled and the number of students taking the final exam) decreased by 15.5% between the traditional lecture (20.5%) and PI (5%) sections at the 2-yr college. Similarly, the implementation of PI at Harvard University reduced the dropout rate by more than half to a rate consistently < 5% (Lasry et al., 2008). Increased retention and lower failure rates in courses with PI have also been reported from a retrospective study of more than 10,000 students in lower- and upper-division computer science courses (Porter et al., 2013).

STAKEHOLDERS’ VIEWS ON PI

What Do Students Think about PI?

Students’ resistance to instructional practices that differ from their expectation (i.e., traditional lecture) has been reported as an important barrier to instructors’ continued implementation of evidence-based instructional practices (Felder and Brent, 1996). Thus, positive student reception of new instructional practices is important. Several studies have investigated this particular aspect of PI. For example, the longitudinal study conducted by Mazur and colleagues found no difference in students’ course evaluations before and after implementation of PI (Crouch and Mazur, 2001). On the other hand, the instructor survey study conducted by Mazur and colleagues found that out of the 384 instructors, 70% reported obtaining higher course evaluations from students in PI classes compared with course evaluations for traditional courses. Despite these overall positive results, 17% of instructors reported a mixed response from students, while 5% reported a negative response. Additionally, a small percent (4%) of instructors who reported that their students had a positive response to PI indicated that the response was initially negative (Fagen, 2003).

In another study, student opinion of PI was compared between majors in a genetics course and nonmajors in an introductory biology course (Crossgrove and Curran, 2008). Each group answered a student opinion survey containing 11 questions. The average Likert scores for all but two questions were not significantly different between the groups. In particular, nonmajors thought that PI improved their exam performance, whereas majors thought this to a lesser extent ($p < 0.001$). The authors also found that the nonmajors were more inclined to encourage the instructors to continue using PI, whereas majors were more ambivalent ($p < 0.05$). Student feedback regarding the continued use of PI in their own and other’s courses has also been explored. For example, in introductory computer science (Simon et al., 2010), exercise science (Cortright et al., 2005), preparatory engineering (Nielsen et al., 2013), engineering mechanics (Boyle and Nicol, 2003), and veterinary physiology courses (Giuliodori et al., 2006), students generally recommend that PI be used in other and/or future courses.
Interestingly, researchers have identified specific aspects of PI that students appreciate. For example, students report that they value the immediate feedback PI provides (Cortright et al., 2005; Giuliodori et al., 2006; Crossgrove and Curran, 2008; Simon et al., 2013a). Moreover, in an analysis of 84 open-ended surveys from students enrolled in an introductory computer science course implementing PI pedagogy, Simon et al. (2013a) found that students felt that PI improved their relationship with their instructors, a finding also observed among students in an exercise science course (Cortright et al., 2005).

Most importantly, students overwhelmingly report that PI helps them learn course material (Cortright et al., 2005; Ghosh and Renna, 2006; Giuliodori et al., 2006; Porter et al., 2011b; Nielsen et al., 2013; Simon et al., 2010, 2013a). Indeed, PI has been shown to significantly impact student self-confidence (Gok, 2012; Zingaro, 2014). For example, students in two sections of an introductory computer science course, one with PI and one with traditional lecture, were asked to rate their self-confidence on a variable of programming tasks (Ramaulingam and Wiedenbeck, 1999) at the beginning and end of the semester (Zingaro, 2014). Students enrolled in the PI course had statistically significant gains ($p = 0.015$) in self-efficacy compared with those enrolled in traditional lecture courses (Zingaro, 2014). In another study, the self-efficacy of students enrolled in algebra-based physics courses with or without PI was compared (Gok, 2012). Similarly, the self-efficacy of the students increased significantly with PI ($p = 0.041$) compared with traditional lecture.

Overall, students have neutral to positive views on PI and seem to recognize its value over traditional teaching.

**What Do Instructors Think about PI?**

In addition to getting students’ opinions, it is also important to gain insight into instructors’ experiences with PI implementation. The instructor survey study conducted by Mazur and colleagues found that 90% of these instructors reported having a positive experience, 79% indicated that they would continue implementing PI, and another 8% reported they would probably use PI again (Fagen et al., 2002). This positive response was echoed in a study conducted by Porter et al. (2011a), who observed that PI was beneficial because it “enables instructors to dynamically adapt class to address student misunderstandings, engages students in exploration and analysis of deep course concepts, and explores arguments through team discussions to build effective, appropriate communication skills” (p. 142). Although a large number of faculty members have reported using PI (Fagen et al., 2002), most research has focused on student perception and learning rather than faculty experience. More information about instructors’ perception of this pedagogy is needed to help inform the successful implementation of PI.

**ACADEMIC SETTINGS IN WHICH PI HAS BEEN IMPLEMENTED**

PI has been implemented in a variety of academic settings (see Supplemental Table 1 for a complete list of studies). For example, out of the 384 PI users surveyed by Mazur and colleagues, 67% taught at universities, 19% taught at 4-yr colleges, 5% taught high school, and smaller percentages taught at other institutions, such as 2-yr and community colleges (Fagen et al., 2002). PI has also been implemented in different subject areas, course levels, and class sizes. In particular, research has been conducted in astronomy, biological sciences, calculus, chemistry, computer science, geosciences, economics, educational psychology, engineering, entomology, medical/veterinary courses, philosophy, and physics courses (see Supplemental Table 1). Of the research articles cited in this paper, 84% used PI in lower-level courses, 12% used it in upper-level undergraduate courses, and 4% used it in medical/veterinary school courses. Furthermore, 25% used PI in small classes (< 50 students), while 75% used it in large classes.

**EVIDENCE-BASED IMPLEMENTATION OF PI**

There are several great resources describing best practices for implementing PI (Mazur, 1997; Crouch and Watkins, 2007). However, these guidelines do not necessarily provide empirical data to support best practices. In the next sections, we describe the results of empirical studies that have tested many of these guidelines. The discussion of guidelines will follow the order of the model for PI presented and researched by Mazur (1997):

1. Question posed
2. Students given time to think
3. Students record individual answers
4. Students convince their neighbors (peer discussion)
5. Students record revised answers
6. Feedback to teacher: tally of answers
7. Instructor’s explanation of correct answer

**Why Does the Type of Question Posed Matter? (Step 1)**

PI is intended to address misconceptions in a specific content area and foster conceptual understanding. To achieve this intended outcome, the type of question asked during each PI event should have an explicit pedagogic purpose (Beatty et al., 2006); however, both the difficulty level of the question and the extent to which instructors ask conceptual questions varies. In an ethnographic study exploring the ways six physics instructors implemented PI, Turpen and Finkelstein (2009) found that the content-related questions instructors asked could be classified as conceptual, recall, or algorithmic. The percent of conceptual questions ranged from 64 to 85%, recall ranged from 4 to 24%, and algorithmic from 0 to 11%. Although instructors primarily asked conceptual questions, recall questions appeared common. Asking recall questions may be appealing to instructors; however, research suggests that asking higher-order questions yields better student results.

In a study in a large medical physiology course for first-year students, Rao and DiCarlo (2000) compared question type with the percentage of correct answers on the individual vote versus after peer discussion. Questions were classified as either testing recall, intellectual (i.e., comprehension, application, and analytical ability), or synthesis and evaluation...
Are Clickers Necessary? (Step 3)

PI is often associated with classroom response systems or clickers. However, not everyone using clickers is conducting PI. Likewise, PI can be conducted without the use of clickers. Several studies have investigated the effect of PI when other voting methods, such as flash cards or ABCD cards have been used. For example, Lasry (2008) compared PI using either flash cards or clickers in two different sections of an algebra-based mechanics course taught by the same instructor. Student learning was measured by comparing learning gains on the FCI and an examination between the two groups. While both the flash cards and clicker groups improved on the FCI, no statistical differences in learning gains were observed between the two groups on either the FCI or the examination. Flash cards have proven effective in other studies. For example, Cortright et al. (2005) studied a physiology course using flash cards and found that students’ problem-solving skills improved during PI.

Another study by Brady et al. (2013) used a quasi-experimental design to investigate differences in students’ metacognitive processes and performance outcomes in sections of a psychology course that implemented PI with clickers versus other sections of the same course that implemented...
PI with paddles (i.e., a low-technology flash card system provided by the primary researcher). Results demonstrate that the use of paddles increased metacognition skills, while the use of clickers resulted in higher performance outcomes. This study indicates that there may be a metacognitive advantage to using non clicker response systems for PI, but additional studies are needed to confirm this finding.

Overall, these studies indicate that PI can be effectively implemented with clickers or with low-tech voting tools.

**Does Showing the Distribution of Answers after the First Vote Matter? (Step 3)**
When personal response devices such as clickers are used, the instructor has the option to share with her/his class the distribution of students’ answers following the first vote. Several studies have examined the impact of this option on students' behavior during peer discussion. Perez et al. (2010) investigated whether showing the distribution of answers (as a bar graph) following the first vote (step 2) biased students’ second vote (step 4). They implemented a crossover research design in a freshman-level biology majors course (eight sections, 629 students participated in the study): in one treatment, students were shown the bar graph before peer discussion, and in the other treatment, they were not. Thus, each treatment group saw the bar graph after the first vote 50% of the time. Instructors in each section used the exact same set of questions. The responses of students seeing the bar graph before peer discussion were compared with those who did not. They found that when students saw the bar graph after the first vote, they were 30% more likely to change their answer to the most common one. This bias was more pronounced on difficult questions, and it appeared to account for 5% of the learning gains observed between the first and second vote. This study suggests that when instructors display the bar graph, students may think that the most common answer is correct rather than constructing a correct response through discussion with their peers. Indeed, data from a qualitative study on the impact of showing the bar graph after the first vote support these findings (Nielsen et al., 2012). Group interviews revealed that students perceive the most common answer to be the most correct, and students are less willing to defend an answer if it is not the most common one.

Student bias toward the most common answer when the distribution of answers is shown was not observed, however, in a similar study in chemistry (Brooks and Koretsky, 2011). Two cohorts of students in a thermodynamics course (n = 128 students) were compared: one saw the distribution after the first vote, while the other did not. They found that both groups of students had a similar tendency to select the consensus answer regardless of seeing the distribution. Moreover, they found no difference in the quality of the explanations students wrote to justify their answers. They did, however, see a difference in students’ confidence in their answers: students who saw the bar graph after the first vote were statistically more confident when their answer matched the consensus answer, even if the consensus answer was incorrect (Brooks and Koretsky, 2011).

More research is needed to fully understand the effect of displaying the bar graph after the first vote. Based on the results of the few studies investigating this issue, it seems that it may be most effective to show the difference in the distribution of answers between the first and second vote after peer discussion. This approach would limit the bias toward the consensus answer observed in some studies (Perez et al., 2010; Nielsen et al., 2012), while not only enhancing the confidence of students who had the correct answer in the first vote but also maintaining the integrity of student discussion.

**When Is It Appropriate to Engage Students in Peer Discussion? (Step 4)**
The analysis from the Why Does the Type of Question Posed Matter? section suggested that the benefits of student discussion on learning vary based on the proportion of correct responses on the initial vote; indeed, limited learning gains between the individual vote and the revote were observed on easy questions (Rao and DiCarlo, 2000; Smith et al., 2009; Porter et al., 2011b; Knight et al., 2013). In their longitudinal analysis, Crouch and Mazur (2001) found that the largest improvement in moving toward the correct answer on a revote occurred when the initial individual answer was correct for ~50% of the class. Nevertheless, there were still substantial learning gains when the initial percent of correct responses was between 35 and 70%. This empirical study indicates that students should be engaged in peer discussion when the percent of correct answers on the individual vote falls between 35 and 70%. Below 35%, the concept may need to be further described or a salient hint provided. Subsequent studies suggest that students may still benefit from talking to one another, even when only a small proportion (~35%) of the class obtained the correct response (Simon et al., 2010; Smith et al., 2009). Above 70%, the instructor should skip to the explanation of the answer.

More research is needed to optimize guidelines for step 4. Regardless of the proportion of correct answers on the initial vote, students seem to benefit from peer discussion.

**Does Peer Discussion Matter? (Steps 4 and 5)**
Peer discussion is the most recognizable feature of the PI model, and much of this review has been devoted to reporting on the learning gains observed after students’ discussions. As such, it is important to understand the role of small group discussion in PI as well as to determine whether or not the observed improvement in student response is more than students with incorrect answers simply copying those who are correct. Smith et al. (2009) investigated this issue in a one-semester undergraduate genetics class (n = 350). During the semester, students were asked 16 sets of paired questions testing the same concepts but with different cover stories. The first question, Q1, was given, and students voted individually, discussed the question with their peers, and voted again (Q1sd). Then, students were given a second question, Q2, testing the same concept as the first. The proportion of correct answers for each question was compared. They found that the proportion of correct answers for Q2 was significantly greater than for Q1 and Q1sd, and out of the students who initially answered Q1 incorrectly but Q1sd correctly, 77% went on to answer Q2 correctly. Thus, when students do not initially understand a concept, they can discuss the material with a peer and then apply this information to answer a similar question correctly. Interestingly, a statistical analysis
of students who answered Q1 incorrectly but Q2 correctly suggests that some students did not belong to a discussion group with a student who knew the correct answer. These students were presumably able to arrive at the correct answer through peer discussion. Porter et al. (2011b) replicated the previous study in two different upper-division computer science classes (n = 96 total). Similarly, they found that more students answered Q2 correctly compared with Q1 and that peer discussion improved learning gains for 13–20% of the students. Peer discussion has also been shown to improve the proportion of correct responses on the re-vote in general chemistry courses (Bruck and Towns, 2009).

Although the proportion of correct answers increases after peer discussion, alternative hypotheses, such as the extra time allowed for individual reflection or to process information, could also explain these learning gains. Lasry et al. (2009) designed a crossover study in three algebra-based introductory physics courses (n = 88) to test whether peer discussion or other metacognitive processes, such as reflection on their own learning process, explained the learning gains associated with PI. Students voted on a question individually and then were assigned one of three tasks: peer discussion, silent reflection on answers, or distraction by cartoon. All groups were asked to vote again. The learning gains were highest when students engaged in peer discussion. These results suggest that the improvement observed after peer discussion is not due to another metacognitive process.

To assess the impact of student discussions, Brooks and Koretsky (2011) examined the relationship between student reasoning before and after peer discussion. Students recorded an explanation for their responses to concept tests before and after discussing them with their peers. Each explanation was then analyzed for both depth and accuracy. The quality of explanations from students who had responded correctly on both the initial vote and the re-vote improved following peer discussion. Even though these students had the correct answer initially, they gained a more in-depth understanding of the concepts after peer discussion. Even though student explanations improve after peer discussion, the actual quality of the discussion appears variable. James and Willoughby (2011) recorded 361 peer discussions from four different sections of an introductory-level astronomy course. When they compared student responses with the recorded conversations, they found that 26% of the time student responses implied understanding, while the quality of conversations suggested otherwise. Furthermore, in 62% of the recorded conversations, student discussions included incorrect ideas or ideas that were unanticipated.

Taken together, these studies suggest that, although peer discussion positively impacts student learning, the improvements observed between the first and second vote may overestimate student understanding.

How Much Time Should Be Given to Students to Enter Their Votes? (Steps 5 and 6)

Once a question is asked, instructors must allow students enough time to respond thoughtfully while still maximizing class time. Faculty members have reported variations in the voting time they allowed students during PI (Turpen and Finkelstein, 2009). For example, during a semester-long observation of six physics instructors, the average voting time given to students varied from 100 ± 5 s to 153 ± 10 s. Out of these six instructors, two had large SDs of more than 100 s. Not only were there variations in response times given between classes, but also within classes (Turpen and Finkelstein, 2009). Thus, establishing guidelines for the optimal amount of time that instructors should allow is important. Unfortunately, we were only able to identify one study on this topic for this review. Miller et al. (2014) examined the difference in response times between correct and incorrect answers in two physics courses with PI: one at Harvard University and one at Queen’s University (Kingston, Ontario, Canada). In both classes, the proportion of correct to incorrect answers decreased when ~80% of the students had responded, suggesting that incorrect answers take more time than correct ones. Additionally, these researchers examined student response times for questions classified as easy or difficult. For easy questions, students who answered incorrectly took significantly more time than those who answered correctly, while for difficult questions, students took more time to answer regardless of correct or incorrect responses.

Although more research is needed to confirm Miller and coworkers’ findings, these data suggest that instructors should alert students that they will terminate polls (i.e., issue a final countdown) after ~80% of students have voted, particularly when posing less difficult questions.

How Does the Role of the Instructor Affect PI? (Step 6)

Although PI’s critical feature is peer discussion, the instructor’s explanation of concept tests also influences the effectiveness of PI. Several studies have investigated the impact of instructors’ explanations at the end of the PI cycle on student learning (Smith et al., 2011; Zingaro and Porter, 2014a,b). Smith et al. (2011) used pairs of isomorphic questions (i.e., two different questions assessing the same concept) to compare the impact of three different instructor interventions in two genetics courses: one for biology majors (n = 150 students) and one for nonmajors (n = 62 students). The three experimental conditions were as follows:

1. Peer discussion only: students answer the first question (Q1) according to the PI model. After the re-vote, the instructor provides the correct answer without explanation, and students answer the isomorphic question (Q2).
2. Instructor explanation only: students answer Q1 individually. The instructor explains the answer to the class, and then students answer Q2.
3. Peer discussion and instructor explanation: students answer Q1 according to the PI model. After the re-vote, the instructor explains the answer, and students answer Q2.

Significantly larger learning gains ( p < 0.05), as measured by the normalized change in scores between the two questions, were observed in the third intervention, which combined PI with instructor explanations (Smith et al., 2011). Moreover, these learning gains were observed across both courses and for students at all levels of ability (low, medium, and high performing as determined by the mean scores on the first question).

An analogous experiment by Zingaro and Porter (2014a) in an introductory computer science class (n = 126) yielded...
similar results. Students experienced larger learning gains with PI through the combination of peer discussion and instructor explanation compared with student discussion alone (81% vs. 69% correct on Q2, respectively). In addition, instructor explanation resulted in the largest gains in learning when the question was more difficult. In a subsequent study, Zingaro et al. (2014b) also found that the combination of peer discussion and instructor explanation compared with peer discussion alone was positively correlated to performance on the final exam.

Instructor behaviors, such as cueing discussion, also appear to impact PI implementation. For instance, Knight et al. (2013) measured the impact of instructional cues on student discussion during PI in an upper-division developmental biology course. In this study, the instructor either framed peer discussion as “answer-centered” (i.e., asking students to discuss answers) or “reasoning-centered” (i.e., asking students to discuss the reasons behind their answers). The resulting student discussions were then analyzed according to a scoring system measuring the quality of student reasoning. The quality of reasoning was significantly ($p < 0.01$) higher in the reasoning-cued condition compared with the answer-cued one (Knight et al., 2013). Other studies have indicated that not only do students perform better on concept tests when given specific guidelines for peer discussion (Lucas, 2009) but also that students place more value on articulating their responses when instructors emphasize “sense making” over “answer making” (Turpen and Finkelstein, 2010).

Together, these data suggest that it is important for instructors to discuss tests with students and to communicate expectations for peer discussion clearly with a focus on sense making.

**Does Grading Matter?**

When using a personal response system during PI, instructors have the option of awarding points for student responses. Awarding points for correct answers (high-stakes grading) versus participation (low-stakes grading) has been shown to impact the dynamics of peer discussion (James, 2006; James et al., 2008; Turpen and Finkelstein, 2010). James (2006) compared student discussion practices in two introductory astronomy classes taught by two different instructors: one a standard class for nonmajors, the other multidisciplinary and focused on space travel. One instructor adopted high-stakes grading, wherein student responses accounted for 12.5% of their total grade. In this class, students were awarded full credit for a correct response but one-third credit for an incorrect one. The second instructor adopted low-stakes grading, wherein student responses accounted for 20% of their total grade, but students were awarded full credit for both correct and incorrect answers. During the semester, 12–14 pairs of student discussions in each class were recorded on three separate occasions. For each conversation pair, James analyzed the conversation bias (i.e., the extent to which one student compared with the other dominated a conversation). The conversation bias among partners was significantly higher ($p = 0.008$) in the classroom with high-stakes grading incentives. Conversation bias was correlated to course grade in the high-stakes classroom but not in the low-stakes one.

In an extension of this research, James et al. (2008) conducted a study examining student discourse in three introductory astronomy classes taught by two different instructors over two semesters. Instructor A taught two semesters, adopting high-stakes grading practices during the first semester and low-stakes grading practices the following semester. Instructor A implemented PI identically each semester and used the same concept test questions. Instructor B adopted a low-stakes grading approach during the first semester but was not observed the following semester. The conversation bias of student pairs was analyzed using two different techniques: one that categorized student ideas (idea count) and another that accounted for the amount of time a student spent talking (word count). When instructor A switched from a high- to low-stakes approach, conversation bias decreased significantly in both the idea count ($p = 0.025$) and the word count ($p = 0.044$) analyses. There was no significant difference in conversation bias between student pairs in either of the low-stakes classrooms taught by the two different instructors. James’ research suggests that when grading incentives favor correct answers, the student with the most knowledge dominates the discussion. Unfortunately, the experimental unit in James’ studies was the students rather than the classroom, which may underestimate the variability of the students, likely confounding these results. However, research by Turpen and Finkelstein (2010) also supports low-stakes grading approaches. For example, they found that high-stakes grading incentives were associated with reduced peer collaboration on questions.

Overall, peer discussion appears to benefit students the most when instructors award participation points for answering questions during peer discussion rather than awarding points for answering questions correctly. It is important to note, however, that although conversation bias improved, more research is needed to link this improvement directly to learning gains.

**LIMITATIONS OF CURRENT RESEARCH ON PI**

Although PI is one of the most researched evidence-based instructional practices, this research has several shortcomings. First, more research is needed to resolve some of the uncertainties in implementation addressed in this review. For example, the impact of high-stakes grading on peer discussion and the consequences of displaying the voting results before peer discussion is unclear. In addition, there is uncertainty regarding the optimal proportion of correct answers needed on the initial vote to engage students in peer discussion. Second, little is understood about the relationship between PI and individual student characteristics or students’ prior knowledge. Although Mazur and colleagues (Lorenzo et al., 2006) reported a reduced gender gap in classes with PI compared with those without, it is unclear whether similar gains occur in other disciplines and whether PI benefits some students more than others. Some research suggests that there is, indeed, a relationship between student characteristics and performance on PI concept tests. For example, Steer et al. (2009) evaluated responses on concept tests from 4700 students enrolled in five different earth science classes at a community college. Female students from underrepresented groups were significantly more likely to change from an incorrect response before peer discussion to a correct one afterward. These students were less likely to have a correct
answer during the individual voting period; therefore, they were able to make larger gains. Similar improvements by underrepresented students have been reported in calculus courses (Miller et al., 2006). Although these results are encouraging, analyzing raw changes in scores does not account for the fact that students with low scores are able to gain more than students who score higher initially. Other studies have shown that gender differences become insignificant once prior knowledge is controlled for (Miller et al., 2014). For example, males and females exhibited significantly different response times on concept tests during both the initial voting period and revote. However, once students’ pre-course knowledge and self-efficacy were accounted for, these differences became insignificant.

Learning gains in PI courses have been analyzed alongside prior achievement in other studies. For example, in a study of 1236 earth science students, prior student achievement (measured by ACT score) predicted the number of correct responses to concept tests during PI (Gray et al., 2011). Unfortunately, overall learning gains were not reported in this study. In another study in computer science courses with and without PI, students’ high school background was compared with scores on a final examination (Simon et al., 2013b). PI was most effective among students indicating that the majority of their high school attended college. For students who indicated that the majority of their high school did not attend college, there was a slight negative, but insignificant, effect of PI versus traditional lecture.

These results imply there is a relationship between the learning gains observed with PI and individual student characteristics such as race, gender, and prior achievement. Thus, the true impact of PI cannot be realized without controlling for student characteristics (Theobald and Freeman, 2014). Researchers must account for the variation between students in classes and between instructors. For example, individual students enrolled in a STEM course for nonmajors in which the instructor has implemented PI should not be directly compared with individual students in another STEM course for majors in which the instructor has not implemented PI. Similarly, differences in instructors’ characteristics, such as fidelity of implementation of PI, demographics, and teaching experience/training must be considered during analysis.

Methodological shortcomings also include inappropriate selection of unit of analysis, which increases the likelihood of type I errors. In particular, this literature commonly reports individuals rather than the classroom in which PI was implemented as the unit of analysis. Individual students within a class are more likely to have similar characteristics beyond those that can be accounted for, and observations from these students are therefore not independent. Choosing the classroom as the unit of analysis ensures independence. Unfortunately, if the unit of analysis is at the aggregate level, studies may be underpowered due to smaller sample size. Thus, it may be useful for researchers to increase the number of classrooms for comparison in their studies or replicate prior studies in order to facilitate meta-analyses.

Finally, the majority of the studies surveyed in this review has utilized classroom response systems, which have grown increasingly more sophisticated. For example, bring-your-own-device technologies are now available and allow instructors to ask open-ended questions. While these new technologies allow instructors to ask higher-level questions, they may unintentionally increase student distraction (Duncan et al., 2012). As institutions and individuals using PI adopt these new technologies, the impact of bring-your-own-device tools on student learning will need to be determined.

CONCLUSIONS

In this review, we provide an analysis of student outcomes associated with PI and features critical to successful implementation. In comparison with traditional lecture, this pedagogy overwhelmingly improves students’ ability to
solve conceptual and quantitative problems and to apply knowledge to novel problems. Students value PI as a useful learning tool and are more likely to persist in courses utilizing it. Likewise, instructors value the improved student engagement and learning observed with PI.

From our analysis of the research, we propose a revised, evidence-based model of the steps of PI (Figure 1). This model could guide practitioners in an effective implementation of PI that would lead to the most positive student outcomes. It could also inform researchers in the design of protocols measuring instructors’ fidelity in PI implementation.

ACKNOWLEDGMENTS

We acknowledge the financial support from the Department of Chemistry and the University of Nebraska–Lincoln. T.V. was funded by the National Science Foundation through grants 1256003 and 1347914.

REFERENCES


President’s Council of Advisors on Science and Technology (2010). Prepare and Inspire: K-12 Education in Science, Technology, Engineering, and Math (STEM) for America’s Future, Washington, DC: Executive Office of the President, President’s Council of Advisors on Science and Technology.


Course-based undergraduate research experiences (CUREs) are being championed as scalable ways of involving undergraduates in science research. Studies of CUREs have shown that participating students achieve many of the same outcomes as students who complete research internships. However, CUREs vary widely in their design and implementation, and aspects of CUREs that are necessary and sufficient to achieve desired student outcomes have not been elucidated. To guide future research aimed at understanding the causal mechanisms underlying CURE efficacy, we used a systems approach to generate pathway models representing hypotheses of how CURE outcomes are achieved. We started by reviewing studies of CUREs and research internships to generate a comprehensive set of outcomes of research experiences, determining the level of evidence supporting each outcome. We then used this body of research and drew from learning theory to hypothesize connections between what students do during CUREs and the outcomes that have the best empirical support. We offer these models as hypotheses for the CURE community to test, revise, elaborate, or refute. We also cite instruments that are ready to use in CURE assessment and note gaps for which instruments need to be developed.
questions for evaluation and research (Westat et al., 2010). However, the simplified lists used in logic models do not capture relationships and feedback loops or the high level of impact certain activities may have. The systems approach to evaluation (Urban and Trochim, 2009; Urban et al., 2014) expands on a traditional logic model by hypothesizing direct and directional relationships among activities and outcomes that evolve over time. Specifically, the systems approach to evaluation does the following:

- Allows for description and visualization of the many diverse connections among activities and outcomes
- Displays multiple pathways students may take to achieve a single outcome
- Accommodates the complexity of education programs by exposing feedback loops and multiple connections among outcomes
- Reveals which outcomes are likely to be most informative by pinpointing where pathways converge

This systems approach to evaluation has been used to design evaluation plans for several existing or planned programs (Urban and Trochim, 2009). We used this approach to generate a broad model of what is known about CURE instruction based on research to date and to identify leverage points at which more information or measures about CUREs are needed. We propose that the resulting model can be used or adapted to generate evaluation plans and an agenda for research on CUREs.

As context for identifying causal mechanisms of CUREs—what makes CUREs work for students—it is useful to consider theoretical perspectives on learning. Social learning theory (Bandura, 1971; Vygotsky, 1978), which has both cognitive and social elements, is particularly useful for considering what makes CUREs beneficial for students. This theory frames learning as a cognitive process, reliant on mental processing and construction, which occurs in a social context through observing and interacting with others. Situated-learning theory is a form of social learning theory that emphasizes the importance of situating learning in an authentic activity, context, and culture (Brown et al., 1989; Lave and Wenger, 1991). For the most part, CUREs are designed to align with the tenets of situated learning. Students do the work that scientists do (e.g., ask questions, design studies, collect and analyze data, build models) in the context of a real scientific problem or question, in which the solution or answer is unknown.

A hallmark of situated learning is “legitimate peripheral participation,” meaning the learner does tasks that experts consider meaningful (“legitimate”) to the work of the discipline. The tasks become increasingly important, moving along a continuum from peripheral to central, as the learner develops expertise (Lave and Wenger, 1991). Students within CUREs are legitimate participants in scientific research, because their actions contribute to achievement of research goals. However, students generally do not perform more central tasks that determine the overall direction and scope of research. For example, in many CUREs, instructors do the central work of posing overarching research questions, which helps steer students in scientifically fruitful directions. Students then do the very real (legitimate) but more peripheral work of collecting and analyzing data to answer those questions. As students develop more expertise, they take on increasingly central roles, for example, by designing or choosing methods and eventually posing research questions themselves.

Working with more expert individuals is another important element of situated learning, because experts model how to do the work, provide feedback about how to improve, and validate the legitimacy of students’ experiences and accomplishments. Situated-learning theory emphasizes the importance of immersion in real culture—in the case of CUREs, the culture of science, which is characterized by scientific ways of thinking, behaving, and working. Through immersion in the culture of science, students not only have opportunities to see science and scientific thinking in action but also to develop in terms of their scientific identity and sense of belonging to the broader scientific community. CUREs are thought to more closely reflect the culture of science through immersion in the process of science when compared with traditional lab-learning experiences (Spell et al., 2014). According to situated-learning theory, we would predict that participating in CUREs would lead to students developing a stronger scientific identity and sense of belonging to the scientific community. We would also expect to observe positive relationships between the degree to which students interact with more expert individuals and the positive outcomes they achieve.

To date, most program evaluation of CUREs has focused on high-stakes outcomes, such as student completion of a science, technology, engineering, and mathematics (STEM) major or matriculation into a science graduate program. Yet the causal elements of these outcomes have not been systematically identified (Corwin Aucinhloss et al., 2014). Although it is important to continue to measure these outcomes, this should be paired with efforts to understand how they are achieved. Building and testing models, such as the ones we present here, will help to direct and improve research and evaluation of CUREs as the biology education community shifts focus beyond “What works?” to “How does it work?” Thus, the purpose of this work is to provide the biology education community with a working model that can be used, adapted, and revised to drive future research on CUREs and assist in CURE evaluation efforts. Our model is intended to show how CUREs “work”—in other words, what it is that students do during CUREs that may lead to the outcomes they experience. To build this model, we:

1. reviewed relevant studies to determine what is known about student outcomes from undergraduate research experiences and how they have been measured;
2. evaluated the studies related to each outcome to determine the extent of empirical support for each outcome—in other words, whether the outcome is probable, possible, or proposed;
3. connected learning activities with the outcomes they are likely to influence to produce several small models of how CUREs work (“mini-models”);  
4. integrated our mini-models into a single, large model to show student progression from activities to short-, medium-, and long-term outcomes; and
5. identified points at which multiple pathways converge or diverge, called “hubs” (see Urban and Trochim, 2009), which we identify as key points for future research and evaluation of CUREs.
METHODS

Step 1. Review of Relevant Literature to Identify Potential CURE Outcomes

We started by assembling a comprehensive set of papers describing studies of CUREs and undergraduate research internships to identify the range of student outcomes. We included studies of internships, because a number of CUREs were developed as a mechanism for involving students in research when sufficient internships were not available (Dolan et al., 2008), and CURE outcomes have been compared with internship outcomes (Lopatto et al., 2008; Shaffer et al., 2010). We included literature cited in a recent report on CURE assessment (Corwin Auchincloss et al., 2014) and in two comprehensive reviews of undergraduate research (Seymour et al., 2004; Laursen et al., 2010). We also searched Google Scholar using the terms “course based undergraduate research experience,” “undergraduate research experience,” “undergraduate research internship,” and “class undergraduate research” for additional studies published within the past 5 yr. All studies included in our analysis met three criteria. They 1) were published within the past 25 yr, 2) presented data gathered in undergraduate settings, and 3) actively examined student outcomes of a CURE or a research internship (i.e., they had elements of experimental design and provided more than anecdotal or descriptive evidence of outcomes).

For the purposes of this paper, we defined a CURE as “a course in which students are expected to engage in science research with the aim of producing results that are of interest to a scientific community.” This definition is purposely broad, since the processes of science research conducted in laboratory courses vary widely, and we have not yet reached consensus on the essential elements that constitute a CURE or make CUREs successful. Laboratory learning experiences described in these studies were deemed CUREs if they reported that the students performed externally relevant scientific research at any point during the course and for any duration. We identified 14 studies on CUREs that met these criteria and 25 studies on undergraduate research internships. In general, the duration of research performed in the identified CURE studies spanned the majority of the course, involved students in multiple science practices (e.g., collecting data, analyzing data, reporting on data), and involved a minimum of 15 students.

We then generated a list of potential CURE outcomes based on evidence or hypotheses presented in this literature sample. We tested the comprehensiveness of this literature sample by searching for additional studies using the same terms described above. We searched two digital libraries (ERIC and JSTOR) and a crowd-sourced database (Mendeley) and reviewed the resulting studies for additional outcomes. We found one new study on CUREs, two studies on internships, and no new outcomes. Thus, we deemed our literature sample sufficient for building models of CUREs. When conducting step 2, we included all studies that fit the four criteria noted earlier.

Step 2. Analysis to Determine the Level of Empirical Support

To determine which outcomes had sufficient empirical support to be included in a model, we designated outcomes as “probable,” “possible,” and “proposed,” using a framework similar to that used by Grayson et al. (2001). Probable outcomes were 1) investigated in a minimum of three studies, 2) measured in at least three different student populations (i.e., groups of students), 3) measured in at least three different courses or curricula, and 4) assessed using at least two different methods or instruments. Possible outcomes were 1) investigated in a minimum of two studies, 2) investigated in two different populations, 3) measured in at least one course or curriculum, and 4) assessed using at least one method. Proposed outcomes were investigated only in a single instance or were supported by learning theory but were not present in the literature sample. Purely descriptive accounts and anecdotes were not used to support these designations. We chose to represent outcomes using discrete ordinal designations in order to determine which outcomes had sufficient support to include in our models. We did not identify any outcomes as certain in our analysis, because current research on CUREs is not sufficient for this designation. We used probable or possible outcomes in subsequent modeling steps.

Step 3. Alignment of Activities and Outcomes to Generate Mini-Models

We connected activities to outcomes to generate several mini-models (Figures 1–3). Following on the systems approach to evaluation (Urban and Trochim, 2009), we parsed outcomes into short, medium, and long term. We defined short-term outcomes as those that can be achieved immediately during a CURE and medium-term outcomes as those that result primarily from achievement of short-term outcomes and occur later, at the end of a CURE or after CURE participation. We defined long-term outcomes as those that result from short- and medium-term outcomes and can only be measured after the CURE is completed. These are often highly valued goals of science education and research experiences in particular, such as development of students’ scientific identity, student persistence in science, and increased public science literacy. All modeling was performed using the Netway tool for program evaluation and planning (Cornell Office of Research and Evaluation [CORE], 2009).

To support the presence and direction of connections in our models, we used social learning theory (described above), investigations of STEM retention outcome relationships (e.g., Estrada et al., 2011), and hypotheses proposed in the CURE literature. We connected activities to short-term outcomes in mini-models and then connected short-to medium-term outcomes. Collectively, the mini-models include the majority of the probable or possible outcomes of CUREs (Table 1 and Figures 1–3).

Step 4. Integration of Mini-Models into a Comprehensive Model

Using the Netway (CORE, 2009), we combined the mini-models into a single large model that depicts the relationships among all activities and outcomes and expands the model to include long-term outcomes (Figure 4). To accomplish this, we used learning theory to hypothesize additional connections among short- and medium-term outcomes and between medium- and long-term outcomes (Brown et al., 1989; Lave and Wenger, 1991).
probable outcomes with the best support were described in the literature. In addition, two studies found that they pursued science graduate degrees or science-related careers at a higher rate than other science majors. A third study showed that CURE students persisted in science majors at a higher rate than students who completed other lab courses. In three additional studies, students reported increased intentions to continue doing research, including pursuing graduate education. This is important, because several researchers have found that students’ educational aspirations are one of the strongest predictors of enrollment in a graduate degree program. We grouped the outcomes of continuing in a science major, entering a graduate program, or pursuing a science career, including students’ intentions to do so and their actual behavior, under the broad heading of “persistence in science.” A number of studies showed evidence of validation of students and their contributions by a broader scientific community, which we call “external validation.” Authorship or acknowledgment in a peer-reviewed journal, presentation at a professional conference, and acceptance of results into a national database all constitute forms of external validation reported in CURE studies.

These results raise several points for consideration. First, probable outcomes with the best support were described

### Table 1. Support for CURE outcomes based on a review of relevant CURE literature

<table>
<thead>
<tr>
<th>Outcome</th>
<th>CURE References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probable</td>
<td></td>
</tr>
<tr>
<td>Increased technical skills</td>
<td>Drew and Triplett, 2008; Shaffer et al. 2010; Jordan et al. 2014; Rowland et al. 2012</td>
</tr>
<tr>
<td>Career clarification</td>
<td>Drew and Triplett, 2008; Harrison et al. 2011; Shaffer et al. 2014</td>
</tr>
<tr>
<td>Possible</td>
<td></td>
</tr>
<tr>
<td>Increased project ownership</td>
<td>Shaffer et al. 2010; Hanauer et al. 2012; Alkaher and Dolan, 2014</td>
</tr>
<tr>
<td>Increased communication skills</td>
<td>Lopatto et al. 2008; Jordan et al. 2014; Alkaher and Dolan, 2014</td>
</tr>
<tr>
<td>Increased motivation in science</td>
<td>Shaffer et al. 2010, 2014; Alkaher and Dolan, 2014</td>
</tr>
<tr>
<td>Increased collaboration skills</td>
<td>Shaffer et al. 2010, 2014</td>
</tr>
<tr>
<td>Increased sense of belonging to a larger community</td>
<td>Jordan et al. 2014; Shaffer et al. 2014</td>
</tr>
<tr>
<td>Increased positive interaction with peers</td>
<td>Shaffer et al. 2010; Alkaher and Dolan, 2014</td>
</tr>
<tr>
<td>Proposed</td>
<td></td>
</tr>
<tr>
<td>Increased access to faculty interaction</td>
<td>Alkaher and Dolan, 2014</td>
</tr>
<tr>
<td>Increased access to mentoring functions</td>
<td>Hanauer et al. 2012</td>
</tr>
<tr>
<td>Enhanced understanding of the nature of science</td>
<td>Russell and Weaver 2011</td>
</tr>
<tr>
<td>Development of self-authorship</td>
<td>Alkaher and Dolan, 2014</td>
</tr>
</tbody>
</table>

*Green shading indicates probable outcomes, yellow shading indicates possible outcomes, and gray shading indicates proposed outcomes.

### Step 5. Identification of “Hubs”

Hubs are points, or “nodes,” in a model at which multiple paths converge (Urban and Trochim, 2009). Hubs are important, because many pathways must pass through a hub for longer-term outcomes to be achieved. Thus, students would be expected to achieve at least one hub outcome in order to achieve longer-term outcomes. For the present study, we designated as hubs those outcomes that had a minimum of six relationships, or directional connections, with other outcomes or activities (Figure 4).

### RESULTS

Our review and analysis revealed eight probable, seven possible, and four proposed outcomes (Table 1). More than half of the outcomes that may result from CURE participation are only possible or proposed in the CURE literature.

### Level of Empirical Support for Outcomes

**Probable Outcomes.** Increases in content knowledge, technical skills, analytical skills, scientific self-efficacy, project ownership, and career clarification are all probable outcomes, according to our criteria. In addition, two studies tracked students after their participation in CUREs and found that they pursued science graduate degrees or science-related careers at a higher rate than other science majors (Bascom-Slack et al. 2012; Hanauer et al. 2012). A third study showed that CURE students persisted in science majors at a higher rate than students who completed other lab courses (Jordan et al. 2014). In three additional studies, students reported increased intentions to continue doing research, including pursuing graduate education. This is important, because several researchers have found that students’ educational aspirations are one of the strongest predictors of enrollment in a graduate degree program (Sewell et al. 1969; Heller, 2001; Mullen et al. 2003; Walpole, 2003; Nevill and Chen, 2007; Eagan et al. 2013). We grouped the outcomes of continuing in a science major, entering a graduate program, or pursuing a science career, including students’ intentions to do so and their actual behavior, under the broad heading of “persistence in science.” A number of studies showed evidence of validation of students and their contributions by a broader scientific community, which we call “external validation.” Authorship or acknowledgment in a peer-reviewed journal, presentation at a professional conference, and acceptance of results into a national database all constitute forms of external validation reported in CURE studies.

These results raise several points for consideration. First, probable outcomes with the best support were described.
in seven unique instances (i.e., different courses or distinct groups of students). Although this met our criteria, future work should focus on investigating the prevalence of these outcomes across a broader set of CUREs that represent the diversity of instruction and students. Second, in a few CURE studies, skill gains were measured through tests, observation, or instructor evaluations of student work. In most, however, students reported gains in their skills. This can be problematic, because novices tend to overestimate their own knowledge and skills, a phenomenon described as the Dunning-Kruger effect (Boud and Falchikov, 1989; Falchikov and Boud, 1989). More direct measurements of technical and analytical skill development in CUREs are needed to corroborate current findings. Third, certain CUREs may be subject to the same recruitment or selection bias as research internships, in that they only involve a small group of high-achieving or research-interested students (Brownell et al., 2013). Moving forward, it will be important to study the impact of CURE instruction using approaches that control for student-level variables and nonrandom assignment (Theobald and Freeman, 2014; Beck and Bliwise, 2014), such as propensity score matching (e.g., Schultz et al., 2011) or regression discontinuity design (e.g., DesJardins et al., 2010, 2014).

Possible Outcomes. We categorized increased communication skills, collaboration skills, motivation to pursue science learning, and enhanced science identity as possible outcomes of CUREs, because they were documented in at least two unique instances (i.e., different courses and different student groups). We classified descriptions of successful peer–peer instruction (Shaffer et al., 2010) and productive scientific discourse with peers (Alkaher and Dolan, 2014) as “positive interaction with peers” and also categorized this as possible. Two studies reported increases in the following two outcomes: 1) tolerance for obstacles and 2) an increased sense of belonging to a larger community (Shaffer et al., 2010; Jordan et al., 2014). Because these studies examined two distinct CUREs, we categorized these outcomes as possible. However, both outcomes were measured with a single question each, using a single instrument for which validity and reliability information is not available. Moving forward, it will be important to establish the validity and reliability of outcome measures and demonstrate outcomes with multiple kinds of evidence and within many different kinds of CUREs.

Proposed Outcomes. In our analysis, we found only single instances of increased access to faculty (Alkaher and Dolan, 2014), increased access to mentoring functions (Hanauer et al., 2012), enhanced access to the nature of science (Russell and Weaver, 2011), and development of self-authorship (Alkaher and Dolan, 2014). These are notable shortcomings, because the development of meaningful relationships with faculty, access to mentoring, and self-authorship are predictors of persistence in science and career decision making, especially for students who are underrepresented in science (Maton et al., 2000; Packard, 2004; Creamer and Laughlin, 2005; Eagan et al., 2010, 2011). More evidence is needed to elucidate whether these outcomes are replicable in particular CUREs or in CUREs in general.

Mini-Models: Connecting Activities to Outcomes
We used both probable and possible outcomes to construct mini-models of connections between activities and outcomes of CUREs. Although there is theoretical support for many of the relationships, the connections we show indicate that most have not been empirically investigated. We intend this set of models to illustrate how to go about pathway modeling of CURE instruction; they are not the only models possible for CUREs or as a comprehensive set of such models.

We chose to focus on six activities that students typically engage in during CURE instruction: reading and evaluating science literature, selecting or designing methods, collecting novel data, analyzing results, working collaboratively, and presenting results outside class (Hathaway et al., 2006; Drew and Triplett, 2008; Caruso et al., 2009; Shaffer et al., 2010; Siritunga et al., 2011; Brownell et al., 2012; Hanauer et al., 2012; Jordan et al., 2014). We do not expect that all CUREs will include all of these activities or be limited to them. If this model is adapted for use in research and evaluation of CUREs, it should be tailored accordingly.

Knowledge and Skills Mini-Model. This model depicts how students develop content knowledge and hone their technical and analytical skills when they read and evaluate scientific literature, collect data, and analyze results (Figure 1; Shaffer et al., 2010; Brownell et al., 2012; Rowland et al., 2012; Siritunga et al., 2011; Kloser et al., 2013). Increases in knowledge and skills lead to improved scientific self-efficacy (Thiry and Laursen, 2011; Hanauer et al., 2012) and, ultimately, increased motivation to learn more science, which further improves science efficacy (Graham et al., 2013).

Communication and Collaboration Mini-Model. This model depicts how, when students work collaboratively...
and communicate about their work outside class (Figure 2; Caruso et al., 2009; Bascom-Slack et al., 2012; Shaffer et al., 2014; Jordan et al., 2014), they improve their communication and collaboration skills (Seymour et al., 2004; Laursen et al., 2010; Jordan et al., 2014; Shaffer et al., 2014). The more students collaborate and communicate, the more they feel they belong to a larger community (Alexander et al., 1998), such as a lab community (Barlow and Villarejo, 2004) or a science learning community (Siritunga et al., 2011; Jordan et al., 2014). Theoretically, students who have opportunities to present to members of a scientific community or who develop professional and personal networks that connect them with a broader community would have increased opportunities for external validation (e.g., publishing; Lave and Wenger, 1991; Wenger, 1998). This external validation further solidifies a student’s role as a community member (Lave and Wenger, 1991). Through feedback, support, and role modeling from more experienced scientists, students can learn to tolerate obstacles and failure and develop the temperament necessary for research work (Thiry and Laursen, 2011; Seymour et al., 2004).

Ownership Mini-Model. This model depicts how students develop a sense of project ownership when they have agency to design their own studies, choose experimental methods, and collect data of interest to them or their community (Figure 3; Hanauer et al., 2012). A growing sense of ownership increases students’ tolerance for obstacles and perseverance, motivating them to complete their projects even in the face of challenges (Ward et al., 2002; Laursen et al., 2010; Hanauer et al., 2012; Alkaher and Dolan, 2014). This has also been conceptualized as “grit” (Duckworth et al., 2007). When students experience success by overcoming obstacles in their research, they develop a greater sense of their scientific self-efficacy, which increases motivation. These outcomes could operate as a feedback loop. When students experience success in overcoming obstacles, their self-efficacy improves, which increases their motivation, which increases likelihood of success in overcoming new obstacles, and so on (Ward et al., 2002, Thiry et al., 2012, Graham et al., 2013).

An Integrated Large Model: Progress toward Long-Term Outcomes

We combined the three mini-models to construct a large CURE model representing our current understanding of how students could achieve outcomes by participating in CUREs (Figure 4). The mini-models are interrelated, because single activities can lead to multiple outcomes and achieving short- and medium-term outcomes is likely to be important for realizing long-term outcomes. To create the large CURE model, we formed two additional connections between short-term outcomes: 1) students who work collaboratively with peers make improvements in their technical skills by receiving modeling and feedback on how to perform particular tasks (Lave and Wenger, 1991); and 2) students who feel ownership over their projects develop an increased sense of belonging to a science community, since they see the science projects as extensions of themselves (Wiley, 2009; Hanauer et al., 2012).

We also formed five new connections between medium- and long-term outcomes and three new connections among
long-term outcomes to form the new pathways depicted in Figure 4. We designated enhanced science identity, career clarification, and persistence in science as long-term outcomes, since they are likely to occur late in CURE participation or after CURE participation. Here, we define “persistence” as students staying in a science track 1 yr after they participate in a CURE. Thus, persistence includes staying in a science major, matriculating into a graduate program, or pursuing a career in science depending on when the CURE falls in a student’s academic trajectory. Many medium-term outcomes have potential to enhance students’ science identity. As students grow in their sense of belonging to a larger scientific or laboratory community and as they receive validation from that community, they are likely to further identify as scientists and decide whether a career in science is of interest (Lave and Wenger, 1991). Tolerance for obstacles is a broadly recognized characteristic of a scientific disposition; self-recognition of this characteristic also enhances science identity (Seymour et al., 2004; Thiry and Laursen, 2011; Thiry et al., 2012). Increased scientific self-efficacy increases motivation and scientific identity, both of which influence persistence in science (Harrison et al., 2011; Estrada et al., 2011; Adedokun et al., 2013).

We expect other factors, such as lower social barriers and increased environmental support (see Lent et al., 1994; Estrada et al., 2011; Graham et al., 2013), to mediate and contribute to identity and persistence in science. Despite support for these hypotheses, they are not modeled, because they have not yet been established as probable or possible CURE outcomes.

---

**Figure 4.** Large CURE model. Arrows represent positive directional relationships between activities and outcomes. Bold black arrows indicate new connections between activities and short-term outcomes in the mini-models. Bold blue arrows indicate new connections between medium- and long-term outcomes.
Identification of Hubs, Reciprocal Relationships, and Feedback Loops

A distinct advantage of the systems approach to evaluation is the process of identifying hubs, reciprocal relationships, and feedback loops. Our model shows that scientific identity is a hub, because multiple activities and short- and medium-term outcomes converge on scientific identity, and scientific identity is on multiple paths to the highly desired outcome of persistence in science (American Association for the Advancement of Science, 2011; President’s Council of Advisors on Science and Technology, 2012). When evaluation resources are limited, evaluation of hubs should be prioritized. For example, if a CURE instructor or designer wants to know whether his or her CURE is effective, it may be best to prioritize measuring changes in students’ scientific identity over changes in content knowledge. Similarly, when resources for instrument development are limited, developing valid and reliable ways of measuring hub outcomes should be prioritized.

In the large model, we identified what we believe to be reciprocal relationships and feedback loops, which we posit may be useful for setting evaluation priorities. We define a “reciprocal relationship” as one in which outcomes are linked via mutual positive relationships. For example, self-efficacy and motivation are likely to be reciprocally linked. As a student develops greater self-efficacy, he or she is likely to be more motivated, which in turn will further his or her sense of self-efficacy. There is good consensus on how to define self-efficacy, and there are published measures of scientific and research self-efficacy (Lent et al., 1994; Chemers et al., 2010; Estrada et al., 2011). A similar consensus does not exist about what constitutes student motivation in science. Thus, priority should be placed on evaluating only one of two reciprocally related activities or outcomes, with a mind to which may be most easily or accurately measured.

We propose that feedback loops occur when an outcome has an indirect effect on itself via downstream outcomes. Feedback loops are distinct from reciprocal relationships, because feedback loops operate over time and should be empirically identifiable. For example, increased self-confidence should lead to increased motivation, which will help students overcome obstacles. This will further increase students’ confidence (Graham et al., 2013). Thus, if students have the opportunity to encounter and fix problems, we should continue to see incremental increases in self-efficacy with each new obstacle they overcome. This can be empirically tested in a longitudinal study. Longitudinal studies will be useful for evaluating longer pathways to achieving outcomes and for identifying low-term feedback loops. For example, designing a project should lead to an increased sense of ownership, which in turn leads to an enhanced sense of belonging to a larger community, enhanced scientific identity, and, ultimately, completion of a science major and possible pursuit of a graduate degree, which will afford more opportunity for identity development and so on. However, this may take years to assess. When time and resources are limited, priority should be placed on measuring short- and medium-term outcomes. If these are not achieved, then effort should be invested in improving CURE instruction to better achieve short- and medium-term outcomes rather than the labor-intensive, long-term tracking of students.

DISCUSSION

The systems approach to evaluation emphasizes the importance of developing a model of a program before selecting which outcomes to measure. We can use models to identify outcomes of particular interest. Urban and Trochim (2009) proposed three categories of outcomes to prioritize: 1) “hubs,” which are highly connected diagnostic outcomes; 2) “low-hanging fruit,” which are easily measurable outcomes; and 3) “pinnacle” outcomes, which are important for stakeholder interests (e.g., funding agencies, institutional priorities, accreditation) or necessary for continuing the program. We should also consider the availability of existing measures (protocols, rubrics, surveys, etc.) in determining which outcomes to assess and when to assess them. We can prioritize early, middle, and late evaluation phases that are developed based on collective consideration of this information, the timeline for CURE development, and the timeline for when students are expected to realize particular outcomes.

We offer three phases for evaluation of CUREs to illustrate this process (Figure 5). We discuss which outcomes constitute low-hanging fruit, pinnacle outcomes, and hubs in each phase. We designed phases around when outcomes are likely to be achieved, taking into account the ease of assessment. We recommend that each CURE evaluation be designed in program-specific phases that consider particular CURE activities, stakeholder interests, programmatic goals, and evaluation resources.

Early-Phase Evaluation

When CURE instructors or designers are first implementing and evaluating their CUREs, this is the place to start. This phase focuses on whether activities are resulting in short-term outcomes. Measures exist for many of these outcomes (low-hanging fruit), including assessment of analytical research skills (e.g., developing hypotheses, designing experiments, analyzing data; Feldon et al., 2011; Sirum and Humbug, 2011; Brownell et al., 2014; Dasgupta et al., 2014; Deane et al., 2014), project ownership (Hanauer and Dolan, 2014), and the hub of self-efficacy (Chemers et al., 2010; Estrada et al., 2011). In addition, these outcomes should be measurable during or immediately after CURE instruction.

Middle-Phase Evaluation

This phase focuses on assessment of the communication and collaboration elements of CUREs and includes one hub: becoming part of a larger community. This phase will require the development and testing of new instruments, because, to the best of our knowledge, there are no measures of communication or collaboration skills or of students’ sense of belonging to a science community that have been tested with CURE students, although there are tools that could be adapted for these purposes. For example, Sevian and Gonsalves (2008) offer a rubric for measuring the effectiveness of scientific explanations, but it was developed for use with graduate students explaining their work to middle-school students. Similarly, there are general measures of undergraduate students’ sense
of belonging or sense of community and the relationship between these and students’ intended or actual persistence in college (Bollen and Hoyle, 1990; Chipuer and Pretty, 1999; Hausmann et al., 2007, 2009). Again, these tools need to be adapted and tested to demonstrate their usefulness for CURE assessment. Some of the outcomes in the middle phase also take longer to achieve. For instance, it may take time to establish a mechanism through which students can present work outside class or it may take multiple iterations of a CURE to yield sufficient data for publication in a science journal.

Late-Phase Evaluation
Late-phase evaluation focuses on medium- and long-term outcomes of CUREs, including scientific identity (Chemers et al., 2010; Estrada et al., 2011), which is both a hub and a low-hanging fruit. As with the middle-phase example, we could not find established, valid, and reliable measures of science student motivation, tolerance of obstacles, or career clarification. There are measures that have some potential to be adapted for CURE assessment. For example, the Science Motivation Questionnaire, developed for use with non science majors, might be adaptable for use with majors, depending on whether and how motivation differs between majors and nonmajors (Glynn et al., 2009). The “grit scale” is a general measure of perseverance and passion for long-term goals that could be adapted to be more science specific (Duckworth et al., 2007). The Undergraduate Research Student Self-Assessment includes a career clarification scale that is well grounded in qualitative research but for which validity and reliability information have not been published (Hunter et al., 2009). Although the “pinnacle” outcome of
persistence in science is the long-term outcome featured here, there are other long-term outcomes likely to be of interest and significant value, such as scientific literacy and understanding of the nature of science.

There is a temptation to focus on assessing pinnacle outcomes because of their value to stakeholders and because they are often more compelling than more immediate outcomes. Yet it is important to keep in mind that pinnacle outcomes, such as persistence in science and understanding the nature of science, may take longer to achieve and may only be achieved for a subset of CUREs or by certain students or following multiple CURE experiences. It is also exceedingly difficult to design and conduct a study that shows a causal relationship between an instructional experience and student outcomes, and this difficulty is compounded when the outcomes take an extended time to realize. For example, we included persistence in science as an outcome with sufficient evidence to include in models used to study CUREs moving forward. However, as noted above, the studies we found did not control for the extent to which students may already be planning to persist in science (self-selection). In addition, these studies did not make use of established methods for designing and interpreting quasi-experimental studies in a way that provides definitive evidence that participation in CUREs improves persistence in science. The CURE community should consider a handful of larger-scale collaborations aimed at examining the relationship between CURE activities and desired long-term outcomes rather than expecting individual instructors to collect data that can be used to answer questions about long-term impacts of CURE instruction.

Each of these models, as well as each pair of nodes and the relationship between them, is a testable hypothesis about how CUREs work. For example, we can investigate whether students who develop an improved sense of self-efficacy are more tolerant of obstacles, whether students who develop technical skills also develop an increased sense of self-efficacy, and whether students who collect novel data improve their technical skills. We expect that other aspects of CUREs not addressed here, such as the duration of the experience, will influence these relationships. It is also important to investigate how student differences, such as gender, major, ethnicity, race, first-generation college-bound status, and other factors known to affect realization of science-related outcomes, may influence students’ experiences with CUREs.

The mechanisms by which these outcomes come about and the connections between activities and outcomes and between short-, medium-, and long-term outcomes have not been empirically investigated. In addition, some outcomes demonstrated in studies of research internships are unexplored or only proposed in CUREs. For example, we found that undergraduates who complete research internships make progress in understanding the nature of science and expand their social and professional networks (e.g., Russell et al., 2007; Adedokun et al., 2012; see Supplemental Table 1), but these outcomes have not been broadly explored in CUREs. Previous work has proposed a set of features or dimensions that distinguish CUREs from other lab-learning experiences (Corwin Auchincloss et al., 2014). Specifically, CUREs are thought to be distinctive because they: 1) involve students in multiple science practices, 2) provide opportunities for students to make discoveries, 3) involve students in work that has relevance outside the classroom, 4) involve students in collaborative work, and 5) provide opportunities for iteration. The activities highlighted here relate to some but not all of these dimensions. Additional models need to be developed and tested to identify what is both necessary and sufficient in the design and implementation of CUREs to make them effective for students.

Another important avenue for study of what makes CUREs effective relates to faculty experiences and outcomes. For example, research and theories of how faculty members decide to change their teaching to more actively engage students and emphasize understanding of science concepts and practices raises interesting questions of the potential for CUREs to be a “gateway” to instructional change. For example, faculty members may be quicker to adopt evidence-based teaching approaches when they teach a CURE because of the CURE’s potential to connect to their research interests and their identity as researchers (Brownell and Tanner, 2012). In addition, CUREs may resonate with faculty members’ own experiences learning to do science, and thus they may be willing to sacrifice “content coverage” to offer course experiences they see as more authentic (Spell et al., 2014).

CONCLUSION

In summary, we identified studies on CUREs and research internships and considered learning theory that is useful for explaining the mechanisms by which students learn through research experiences. Then, we characterized the outcomes that students have the potential to realize from participating in CUREs. Using a subset of outcomes that had the most empirical support, we used a pathway-modeling process to generate models of the components and contexts that make CUREs effective for students. We offer these models as hypotheses for the CURE community to test, revise, elaborate, or refute. We cite instruments that are ready to use in CURE assessment and note gaps for which instruments need to be developed. We hope that the community effort to assess CUREs using a common set of tools, which began with the efforts of Lopatto and colleagues (Lopatto, 2004, 2007, 2008; Lopatto et al., 2008) will embrace this new set of tools such that we can continue to compare results across CUREs and between CUREs and research internships. We describe our approach so others can use it to build and test their own models of CURE instruction, which is especially important given the diverse long-term outcomes that may result from research experiences. Understanding how CUREs function will help improve existing CUREs, aid in the design of new CUREs, and promote a common understanding of the utility of CUREs as an educational intervention.

ACKNOWLEDGMENTS

Thanks to Melissa Aikens and Lucas Wachsmuth for their careful reading and thoughtful feedback on drafts of the manuscript and to the Cornell Office for Research on Evaluation for early use of the Netway (CORE, 2009) to develop pathway models. Support for this work was provided by a grant from the National Science Foundation (NSF DBI-1061874). M.J.G. is supported through a Howard Hughes Medical Institute’s (HHMI) professor grant (originally to Jo Handelsman), as well as by an NSF grant (NSF-TUES 1323258). The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of the HHMI or the NSF.
REFERENCES


Urban JB, Hargraves M, Trochim WM (2014). Evolutionary evaluation: implications for evaluators, researchers, practitioners,
funders and the evidence-based program mandate. Eval Prog Plan 45, 127–139.


Considering Student Voices: Examining the Experiences of Underrepresented Students in Intervention Programs

Gina Sanchez Gibau

Department of Anthropology, Indiana University–Purdue University Indianapolis, Indianapolis, IN 46202

Submitted June 23, 2014; Revised April 23, 2015; Accepted May 4, 2015

Monitoring Editor: Deborah Allen

Qualitative studies that examine the experiences of underrepresented minority students in science, technology, engineering, and mathematics fields are comparatively few. This study explores the self-reported experiences of underrepresented graduate students in the biomedical sciences of a large, midwestern, urban university. Document analysis of interview transcripts from program evaluations capture firsthand accounts of student experiences and reveal the need for a critical examination of current intervention programs designed to reverse the trend of underrepresentation in the biomedical sciences. Findings point to themes aligned around the benefits and challenges of program components, issues of social adjustment, the utility of supportive relationships, and environmental impacts.

INTRODUCTION

For the past three decades, the scientific community has expressed great concern about the apparent disparity in participation among historically underrepresented groups in science education and in the sciences in general. This underrepresentation continues to persist, despite the creation and implementation of interventions designed to reverse this trend and ensure the development of a strong future workforce. Some of the more successful strategies that have emerged have enabled underrepresented minority (URM) students to gain greater access to institutions in the fields of science, technology, engineering, and math (STEM) by virtue of increased funding opportunities. Other strategies have focused on the concomitant challenge of retaining and graduating URM students in the sciences to meet the growing demand for such professionals. To this end, research has revealed the benefits of mentoring, professional development, socializing experiences, and the cultivation of inclusive campus environments for these students. Yet underrepresentation continues to exist, often leaving the academic and scientific communities at a loss for a solution to this dilemma.

This study contributes to existing literature on effective approaches, models, and interventions to address the chronic condition of underrepresentation in the biomedical sciences. Most of the research on the participation of URM students in the sciences has been quantitative or mixed method in nature, which is perhaps understandable, given the background in scientific inquiry of researchers within STEM fields. Such studies often deploy surveys and other assessment instruments to deduce the relative success of URM students in STEM disciplines, focusing on such aspects as motivation, persistence, self-efficacy, and the impact of climate and mentoring, to name only a few (Hung et al., 2007; Chang et al., 2008; Griffith, 2010; Estrada et al., 2011; Coronado et al., 2012; Hernandez et al., 2013; Kendricks et al., 2013; Thompson and Campbell, 2013; Salto et al., 2014). While such analyses are critically important to understanding the dearth of participation, they may not explain the totality of URM student experiences. Qualitative studies that offer the richness and depth of the voices and perspectives of URM students themselves are comparatively few (Lewis and Collins, 2001; Gardner, 2008; Hurtado et al., 2009; Johnson et al., 2011; Palmer et al., 2011; Dickins et al., 2013; Prunuske et al., 2013). This study uses a phenomenological approach to examine the experiences and perceptions of URM graduate students...
students in the biomedical sciences of a large, midwestern, urban university. The perspectives offered by URM students, as captured in transcripts from 24 individual interviews and one focus-group interview centering on the evaluation and revision of existing intervention programs in which the students participate, provide a critical contribution to ongoing conversations regarding underrepresentation in the sciences.

Considering the voices of underrepresented students themselves in the design and implementation of these programs is an intervention in and of itself. Taking a phenomenological approach to understanding the experiences of underrepresented students in the sciences is a necessary means for close examination of what types of interventions may or may not work at a given institution. Qualitative studies are particularly valuable, since they offer a more nuanced viewpoint on assessing the efficacy of such interventions and serve to complement existing quantitative research. Likewise, they can serve to enhance and enrich program conceptualization, design, implementation, and revision.

The purpose of the current study was to examine the perspectives and experiences of URM graduate students engaged in National Institutes of Health (NIH)-funded intervention programs and to assess program effectiveness from the students’ accounts. Questions that served as a guide to this study include the following: What were the experiences of the URM students? Which experiences were most critical to their persistence in their programs? And how can these experiences potentially inform and enhance intervention strategies?

Broadening Participation
Much of the literature on historical underrepresentation in the sciences focuses on both the source of the underrepresentation and possible solutions. Aside from the now more recognized barriers to broadening participation, such as lack of funding, mentors, support, preparation, and professional development opportunities (Castle, 1993; Davidson and Foster-Johnson, 2001; Summers and Hrabowski, 2006; Powell, 2007; Butts et al., 2012), some researchers cite increases in retirement that occur over time as contributing to underrepresentation, given the lack of a robust and diverse pipeline of individuals who are not yet ready to enter the workforce (Olson, 1988; National Science Board, 2004). Others express the concomitant concern of increased competition in the U.S. job market due to the steady influx of internationally trained scientists who dominate the current workforce (Campbell, 2000; Hurtado et al., 2009; National Academy of Sciences, 2011). Still others identify a need for URM students to pursue careers as research faculty members, so they can eventually serve as role models for an increasing URM college population (Nelson, 2007). Indeed, URM attrition at the undergraduate level also contributes to the lack of depth of a pool of applicants for the PhD (Russell and Atwater, 2005). Finally, lack of social integration in the culture of science and nonacademic commitments (e.g., economic, familial) are often cited as contributing factors (Davis et al., 2004).

To address the persistence of underrepresentation in the sciences, researchers have been compelled to examine interventions designed to increase a variety of positive outcomes pertaining to the successful achievement of the PhD (Fagen and Labov, 2007). Several factors have been identified that pertain to realms of support, including social, familial, academic, financial, and institutional. Drilling down to the critical components, these would include the following:

- Mentorship, with the quality of the mentoring relationship identified as a key factor
- Institutional climate, in which the institution embodies the value of diversity as stated in its mission
- Funding, with targeted grant programs being the most successful
- Research opportunities, especially those opportunities occurring during the summer months, which then inevitably enhance students’ studies and lab work conducted during regular semesters
- Social integration, which is of particular concern for URM students at predominantly white institutions (PWIs) and critical to the development of a sense of belonging to a larger community of scientists and to students’ acquisition of an identity as a scientist
- Critical mass of URM students and faculty members, which must also be buttressed by a high value for diversity
- Supplemental/flexible instruction or curricula, with which URM students have excelled and which speaks to a more individualized manner for mentoring a student

While the literature pertaining to the experiences of URM doctoral students is growing, this review will focus on the literature that is most germane to this study, pertaining to mentorship, science identity, and science enculturation in general.

The Mentor–Protégé Relationship
The mentor–protégé relationship is critical to the academic success and career advancement of any student pursuing a degree in the sciences (National Science Foundation, 1991; Prunuske et al., 2013). In identifying key indicators for Hispanic doctoral student success, Millett and Nettles (2006) highlight the importance of research assistantships, indicating that “research assistantships may influence other doctoral student experiences such as student interactions with peers, their academic interactions with faculty, their interactions with their faculty advisor, whether students stop out of their doctoral programs, and their rate of progress” (p. 271). A successful doctoral student is one who will have a substantial amount of publications written and who will have had the opportunity to engage in other professional development activities, such as disseminating research at national conferences (Millett and Nettles, 2006). If the mentor–protégé relationship is strained, there are tremendous ramifications for the student (Prunuske et al., 2013). One bad experience in the lab, for example, a protégé being asked to leave or fired, can reverberate throughout the department and create a negative impression on all other preceptors with whom the student could have worked.

Turner and Thompson (1993) examine the mentor–protégé relationship as it pertains to doctoral students who are minority and female. Their research raises the question of whether or not minority women receive adequate opportunities for graduate school socialization in their programs compared with their majority colleagues, bearing in mind that the culture into which students are socialized has been
historically and predominantly white and male (Turner and Thompson, 1993, p. 357; Johnson et al., 2011). This question is gauged on a number of factors, including university recruitment, opportunities for mentorship, perception of departmental environment (cooperative or competitive), and individual experiences with racial and gender discrimination. The study found that most minority female doctoral students experienced less recruitment and fewer apprenticeship opportunities than their majority colleagues. The apprenticeship opportunities were measured by coauthorship, copresentations, referrals to job searches, teaching assistantships, and mentorship from one’s graduate advisor. These students also experienced their departmental environment as competitive, in contrast to majority students, who described their environment as cooperative (Turner and Thompson, 1993, p. 362).

Overall, the minority female doctoral students encountered fewer socializing experiences, which is troubling, given the importance of the apprenticeship experience to the successful completion of the graduate degree and the eventual securing of a career in academia. Enculturation or socialization into the community of science is critical to URM students’ successful acquisition of their identity as scientists, which is a necessary prerequisite to navigating the culture of science, both during their training and afterward when pursuing their career trajectory (Gardner, 2008).

**Scientist Identity Development**

In their study of the impact that race and gender have on science identities, Carlone and Johnson (2007) utilize a science identity model, which includes indicators of competence, performance, and recognition, upon which students must rate themselves and be rated by others as a “science person” in order to demonstrate a strong science identity. According to Carlone and Johnson (2007, p. 1192), a science identity is “accessible when, as a result of an individual’s competence and performance, she is recognized by meaningful others, people whose acceptance of her matters to her, as a science person.” They further define the science identity by the following criteria:

She is competent; she demonstrates meaningful knowledge and understanding of science content and is motivated to understand the world scientifically. She also has the requisite skills to perform for others her competence with scientific practices (e.g., uses of scientific tools, fluency with all forms of scientific talk and ways of acting, and interacting in various formal and informal scientific settings). Further, she recognizes herself, and gets recognized by others, as a “science person.” (p. 1190)

Studies have shown how URM students who have acquired an identity as a scientist fared better than those who did not (Johnson et al., 2011). Conversely, those who experienced instances in which their identity as a scientist was doubted by the institutional support system found it more difficult to navigate the culture of science based on their gender, race, and class status. This navigation process was necessary, since the experience of being doubted effectively “interrupted” their goals of becoming a scientist (Carlone and Johnson, 2007). These “interrupted” students “expressed dissatisfaction about how they were positioned in science and felt their goals to become scientists and doctors were disrupted” (Carlone and Johnson, 2007, p. 1197). The disruption also occurred when students acquired information regarding their professors’ perceptions of them as students. These perceptions included whether or not the professor believed the students to have the requisite skills and ability to succeed on the one hand, versus the perception of them as “exceptional” students if they performed particularly well on the other hand. Students who encountered a disruption in the development of their science identities often “felt overlooked, neglected, or discriminated against by meaningful others within science,” namely their mentors (Carlone and Johnson, 2007, p. 1202).

Carlone and Johnson (2007, p. 1207) concluded that women of color pursuing science as a career path experience one of three phenomena: 1) they are disrupted briefly in their recognition as science people; 2) they seek recognition from relationships outside the traditional mentor–protégé model; or 3) they redefine the parameters of recognition for themselves, in order to persist in their programs and to understand their presence as women of color within a university science community. Those who did persist in the sciences despite the disruption often found ways to negotiate their educational experiences and eventually came to see themselves as scientists (Carlone and Johnson, 2007; Johnson et al., 2011). This negotiation process is a pivotal skill for URM students to acquire if they are to effectively navigate the cultural environments in which science education takes place. It also relates back to mentorship, as one must have a mentor who not only effectively guides the student through the program but also aids in the socialization process to ensure the student acquires the identity of a scientist.

**The Culture of Science**

Female and URM students in the biomedical sciences face the challenge of confronting the tradition of the culture of science, one often associated with masculinity and a historical Eurocentric base (Brickhouse, 2001; Johnson et al., 2011). Carlone and Johnson (2007, p. 1207) caution, “Recognition can thus be viewed as a mechanism for reproducing the status quo in science. It is much easier to get recognized as a scientist if your ways of talking, looking, acting, and interacting align with historical and prototypical notions of scientist [traditionally, white and male].” The current presence of female and URM students and students with disabilities in institutions challenges this long-standing culture, which must undergo change and adaptation to accommodate an agenda of broadening participation of diverse individuals engaged in science education.

Research focusing on the agency of women of color, in particular in negotiating their identities as scientists, illustrates the challenges faced in the educational process (Johnson et al., 2011). Some of these students have encountered incongruence with the science curriculum. The challenge for these students lies in the enculturation process; URM students, particularly those who may come from historically black colleges and universities (HBCUs) or other minority-serving institutions (MSIs), are asked to immerse themselves in a culture of science and acquire identities as scientists, but in order to do so, they are asked to perform in ways to which...
they are not accustomed or to adjust to different pedagogical approaches with relative ease (Stassun et al., 2008; Hurtado et al., 2009). Students in the Carlone and Johnson (2007) study had to make a shift in their course performance, from simply doing well on exams to applying their knowledge and thinking like a scientist.

The ability to think like a scientist and to acquire an identity as a scientist is paramount to URM students’ successful socialization into and completion of graduate degree programs. Ultimately, any intervention created to address underrepresentation in the sciences must consider these factors.

METHODS
The methodological approach to this study was phenomenological by design (Patton, 1990; Creswell, 2009). Narrative data were extracted by document analysis, which entailed reviewing, open coding, and analyzing interview data from program participants over a 6-yr period, as transcribed by the program external reviewer. With a qualitative methodology, document analysis enables the researcher to make meaning out of student experiences based on what was reported (Patton, 1990). A phenomenological approach, in particular, allows the researcher to uncover core meanings by investigating commonalities in experiences among study participants, in this case among URM students in a particular intervention program. Student narratives bring to light the lived experiences of minority students as they progress through their degree program. The idea is that students, speaking for themselves, are far more expert in their own cultural ways of being and thus are equally suited to comment on a particular program and offer suggestions for change.

Program Context
The interviews from which data were extracted were conducted as part of an annual evaluation of two NIH-funded intervention programs at a large, “high research activity,” midwestern, urban university: Bridges to the Doctorate; and the Initiative for Maximizing Student Diversity (IMSD). The intended goal of the Bridges program was to increase the number of URM students who matriculate into and graduate from PhD programs in the biomedical sciences at the medical school or any other doctoral-granting institution upon completion of their master’s degrees at a partnering MSI. In this study, the “bridge” spanned the master’s program at an HBCU in the South and the doctoral program in the medical school at the midwestern urban university. Partnering with an HBCU was critical, since these institutions produce a disproportionate number of URM students who earn bachelor’s degrees in the sciences (Leggon and Pearson, 1997) and who go on to pursue doctorates (Solorzano, 1995; Syverson and Bagley, 1999).

The Bridges program, initially funded in 2003, consisted of the HBCU master’s students spending one to two summers at the midwestern urban institution conducting research with a mentor, with the hope that the students would then leverage this experience toward the successful completion of their degrees at their home institution. The program provided students with a 12-mo stipend and fee remission for those who had the opportunity to engage in regular semester research as well.

In addition to financial support, the program offered students multitiered mentoring through interaction with research faculty members, postdoctoral students, other graduate students, lab personnel, an academic advisor, and advisory board members (Gibau et al., 2010). Through the Bridges program, URM students had the opportunity during the summer to acquire a mentor at their host institution who could thereafter be engaged along with faculty members from the HBCU in monitoring their progress toward the master’s degree. The hope, thereafter, would be that the students would pursue their doctorates at the host institution or a similar institution (Gibau et al., 2010).

In contrast, the IMSD program, initially funded in 2007, provided 2 yr of graduate school funding for URM students in the 10 biomedical sciences PhD programs at the midwestern medical school. In addition, it provided funding for some laboratory supplies and the opportunity for students to present their research at a national meeting. Ostensibly, the IMSD program was created to meet the needs of Bridges alumni who would transition into the PhD programs in the medical school. This grant provided support for the first 2 yr of a student’s doctoral work, with the student garnering sufficient mentoring to apply for and secure external funding for his or her remaining years in the doctoral program.

The Bridges program began with a small cohort of three students in 2003 and grew to 27 individuals as of Fall 2013. In 2010, program coordinators expanded the program and initiated a second partnership with a Hispanic-serving institution located on the West Coast and a third with another southern HBCU in 2013. Of the total 18 Bridges students tracked over time during the course of this study, only three dropped out of the program entirely, while 12 eventually matriculated into doctoral programs.1 Six of these 12 individuals became IMSD-funded students at the midwestern urban institution.

Participants
The participants whose narrative responses are captured in the data set were accepted into the Bridges program and were currently enrolled as graduate students, either as master’s students at the MSI or as PhD students at the partnering institution, between 2003 and 2008. Eighteen graduate students engaged in annual interviews with an external reviewer during this time period. Of the 18 participants interviewed, 16 were women and two were men.2 All of the participants were African American.

Data Collection and Analysis
The data collected and used in this study were part of an ongoing exempt study conducted as part of regular program review (IRB EX0610-29). A purposeful sample of transcripts from 24 individual, standardized, open-ended interviews

1Of the remaining three individuals, two pursued careers outside medicine, and a third obtained a terminal master’s degree from the medical school.
2The apparent gender inequity may speak to the contemporary distribution of women to men in all areas of higher education.
and one focus-group interview conducted by the external reviewer over a 6-yr period was selected and independently analyzed (see Table 1). The transcripts were divided into cohorts determined by the participants’ year of entry into the Bridges program. A total of six cohorts were delineated (see Figure 1). Because the purpose of the study was to determine the meaning that students attributed to their experiences as participants of the grant programs, responses on the interview transcripts were first open coded and then reread, focus coded, and interpreted. Units of meaning were determined inductively and then grouped into patterns of themes, which were interpreted. Thematic analysis allowed the data to be further divided into structures based on cohorts of students as a means of delineating patterned evidence.

RESULTS

Analysis and interpretation of the interview transcripts revealed several themes that characterized the experiences of the URM graduate students participating in the programs, clustered around students’ perceptions of program benefits and challenges. The benefits included the summer research experiences, mentorship, social and professional support, and career definition. The students also encountered specific challenges, including environmental and curricular adjustment as well as status differentiation.

Summer Research Experience

A key theme that emerged from the interview data were the benefits received through summer research experiences. Not
Mentoring Relationships

Another benefit for students were the mentors who expressed a vested interest in ensuring their success from the outset. Students were matched with mentors based on their general interest in a particular research area or based on the availability of space in a particular lab. Some mentors chose to contact their assigned students before their arrival at the partnering institution; these mentors would often send articles for the students to read, to prepare them for what they would encounter in terms of the mentors’ research agendas. This type of outreach relates back to the literature on socialization, with the mentor taking an early and proactive means through which to guide students through their transition.

The positive outcome of summer research experience is aligned with what has been documented in the literature on effective interventions, with the contrasts in access and scale of resources most evident.

The relatively positive comments provided by the students about their mentors speak not only to their assessments of the quality of their mentoring experiences but also to the type of mentorship that is most desirable:

“Others experienced mentors who would meet with them weekly or twice a month. Many times, the mentors would meet with the entire team of students and researchers; the students later pointed out that there was some desire to have alternative methods, like one-on-one meetings. But none of the students indicated any difficulties with gaining access to their mentors or any relationship problems with their mentors. This is key to the development of what was mentioned in the literature as the need for strong and high-quality mentor–protégé relationships to ensure student success. For the most part, they mentioned how their mentors were always willing to help and would encourage the students to seek them out if they were encountering any difficulty:

“So the mentor, she was just great and very helpful. Even before I started the rotation, she would send me articles and different things, like if she had a speaker coming in or a conference that they had on campus.… I thought that was great that she took an interest early on, even before I rotated through her lab. That was cool.”

Despite these comments, there were others that indicated the mentor was not the first person sought out when encountering difficulties in the lab or the classroom. Interestingly, many of the students interviewed talked more about the help they received in the lab from the staff than from the mentor. Many times, the student, if in the lab, would obviously go to the personnel in their immediate vicinity, the postdocs and PhD students. In these instances, the student would receive immediate attention. Thus, many students credited the PhD students and the postdocs with their successes in the lab:

“They [the postdocs] were great. I don’t think I could have done it without them, because … like the techniques and the different procedures and equipment. They helped me out a lot. They helped me make my own protocols and things to follow, and it let me use their protocols, and explained … they didn’t just tell me how … this is what you do. They explained why you do it, and the whole process. It really helped me understand exactly why I was doing what I was doing.”
Access to minority role models through programmed lectures and luncheons was another evident benefit. Yet some students also questioned this strategy based on who was invited to visit:

“The strengths are it gives you an opportunity to, expose you to areas of research, and, that you never would have been able to do at [the HBCU]. It also exposes you to mentors who are excellent teachers and who can give you really good advice. It allows you to attend scientific conferences. I never would have gotten a chance to present at national meetings.”

“It’s basically done everything for my career. Before this, I had no, … before Bridges I really didn’t know what I was going to do once I finished my undergraduate. I was kind of … I had applied to the graduate school at [the HBCU], but not really knowing what I was going to do. But with Bridges, I was able to find direction.”

“At first I was just wanted to, after my master’s or maybe even after my undergrad, I kind of wanted to teach high school or something. Bridges helped to motivate me to go further with my degree, just wanting to do more as far as research goes.”

As mentioned earlier, the literature supports the idea that professional development opportunities, those connected to one’s research and inclusive of the ability to interact with scholars at national meetings, are a critical indicator of the likelihood of students pursuing science careers. Effective mentorship is critical to this process of career definition for the students.

Environmental Adjustment

Students in this study also identified significant challenges to their success and persistence in the program based

These narratives relate back to the idea of students needing the recognition of “meaningful others” as a “science person.”

In these instances, the students’ peers became those meaningful others; they held each other accountable for their mutual success in the program and thus fostered a sense of collective responsibility in their acquisition of their science identities.

Indeed, what stands out most strongly in these narratives are the ways in which the students relied upon one another to “make it” in the program. The program was designed to offer the students opportunities for cohort building and socialization experiences. During their summer experience at the midwestern medical school, they were often housed together in a dorm or apartment. The Bridges and the IMSD students were also expected to attend grant-sponsored luncheons with committee members or with the medical school speaker brought to town, as mentioned earlier. In addition, the students often got together at local restaurants or one another’s apartments, most likely on the weekends. So the students had the ability to see one another on multiple occasions outside the confines of the medical school and share their daily experiences. Whether these experiences had to do with receiving a particular grade on an exam or working through a particularly challenging lab experiment, the students found in one another a sympathetic ear, one that served as a means to validate their presence and thus motivate them to persist in the program.

Career Definition

Finally, the summer research experience also aided student with career definition. The program exposed students to the possibilities of pursuing a career in academe, an alternative to the lure of industry. Much of this exposure came from their interactions with their mentors, their lab work, and the professional development opportunities afforded to them:

In these instances, the students’ peers became those meaningful others; they held each other accountable for their mutual success in the program and thus fostered a sense of collective responsibility in their acquisition of their science identities.

Indeed, what stands out most strongly in these narratives are the ways in which the students relied upon one another to “make it” in the program. The program was designed to offer the students opportunities for cohort building and socialization experiences. During their summer experience at the midwestern medical school, they were often housed together in a dorm or apartment. The Bridges and the IMSD students were also expected to attend grant-sponsored luncheons with committee members or with the medical school speaker brought to town, as mentioned earlier. In addition, the students often got together at local restaurants or one another’s apartments, most likely on the weekends. So the students had the ability to see one another on multiple occasions outside the confines of the medical school and share their daily experiences. Whether these experiences had to do with receiving a particular grade on an exam or working through a particularly challenging lab experiment, the students found in one another a sympathetic ear, one that served as a means to validate their presence and thus motivate them to persist in the program.

Career Definition

Finally, the summer research experience also aided student with career definition. The program exposed students to the possibilities of pursuing a career in academe, an alternative to the lure of industry. Much of this exposure came from their interactions with their mentors, their lab work, and the professional development opportunities afforded to them:

The strengths are it gives you an opportunity to, expose you to areas of research, and, that you never would have been able to do at [the HBCU]. It also exposes you to mentors who are excellent teachers and who can give you really good advice. It allows you to attend scientific conferences. I never would have gotten a chance to present at national meetings.”

“It’s basically done everything for my career. Before this, I had no, … before Bridges I really didn’t know what I was going to do once I finished my undergraduate. I was kind of … I had applied to the graduate school at [the HBCU], but not really knowing what I was going to do. But with Bridges, I was able to find direction.”

“At first I was just wanted to, after my master’s or maybe even after my undergrad, I kind of wanted to teach high school or something. Bridges helped to motivate me to go further with my degree, just wanting to do more as far as research goes.”

As mentioned earlier, the literature supports the idea that professional development opportunities, those connected to one’s research and inclusive of the ability to interact with scholars at national meetings, are a critical indicator of the likelihood of students pursuing science careers. Effective mentorship is critical to this process of career definition for the students.

Environmental Adjustment

Students in this study also identified significant challenges to their success and persistence in the program based

These narratives relate back to the idea of students needing the recognition of “meaningful others” as a “science person.”

In these instances, the students’ peers became those meaningful others; they held each other accountable for their mutual success in the program and thus fostered a sense of collective responsibility in their acquisition of their science identities.

Indeed, what stands out most strongly in these narratives are the ways in which the students relied upon one another to “make it” in the program. The program was designed to offer the students opportunities for cohort building and socialization experiences. During their summer experience at the midwestern medical school, they were often housed together in a dorm or apartment. The Bridges and the IMSD students were also expected to attend grant-sponsored luncheons with committee members or with the medical school speaker brought to town, as mentioned earlier. In addition, the students often got together at local restaurants or one another’s apartments, most likely on the weekends. So the students had the ability to see one another on multiple occasions outside the confines of the medical school and share their daily experiences. Whether these experiences had to do with receiving a particular grade on an exam or working through a particularly challenging lab experiment, the students found in one another a sympathetic ear, one that served as a means to validate their presence and thus motivate them to persist in the program.

Career Definition

Finally, the summer research experience also aided student with career definition. The program exposed students to the possibilities of pursuing a career in academe, an alternative to the lure of industry. Much of this exposure came from their interactions with their mentors, their lab work, and the professional development opportunities afforded to them:

The strengths are it gives you an opportunity to, expose you to areas of research, and, that you never would have been able to do at [the HBCU]. It also exposes you to mentors who are excellent teachers and who can give you really good advice. It allows you to attend scientific conferences. I never would have gotten a chance to present at national meetings.”

“It’s basically done everything for my career. Before this, I had no, … before Bridges I really didn’t know what I was going to do once I finished my undergraduate. I was kind of … I had applied to the graduate school at [the HBCU], but not really knowing what I was going to do. But with Bridges, I was able to find direction.”

“At first I was just wanted to, after my master’s or maybe even after my undergrad, I kind of wanted to teach high school or something. Bridges helped to motivate me to go further with my degree, just wanting to do more as far as research goes.”

As mentioned earlier, the literature supports the idea that professional development opportunities, those connected to one’s research and inclusive of the ability to interact with scholars at national meetings, are a critical indicator of the likelihood of students pursuing science careers. Effective mentorship is critical to this process of career definition for the students.

Environmental Adjustment

Students in this study also identified significant challenges to their success and persistence in the program based

These narratives relate back to the idea of students needing the recognition of “meaningful others” as a “science person.”
on their experiences. One concerns the mere adjustment to social and cultural climate, as experienced at two different types of institutions:

“There is an adjustment to two vastly different environments. You’re coming from [the HBCU], coming up here. It is somewhat of an adjustment, so maybe you might need to brace for kind of that shock a little bit.”

“It’s different. Because I came from two HBCUs and so it is a totally different environment as far as that’s concerned. We use to go around [laughs] at my old school and be in awe when we saw white people in the hall.”

“Yeah, just walking past they say, ‘How are you doing?’ When we first got here, especially me, but I know all us, we were walking past and we were like ‘Oh, hi!’ If I make eye contact with you I think we are supposed to speak, and people are like ‘I don’t know you, why are you talking to me’?”

The latter statement speaks to the difference between regional norms of greeting. The impact of the shift in environment from an HBCU to a PWI for URM students is one that should not be overlooked. Indeed, a holistic approach to student success requires a consideration of their learning environments. In other words, acknowledging and attending to the realities of cultural adjustment experienced by students coming from the supportive culture of an HBCU into an overtly competitive one at a Research I institution, inclusive of an understanding of the time actually needed to make such an adjustment, should be considered in the design and implementation of an intervention strategy (Stassun et al., 2008).

Curricular Adjustment
Related to this shift from HBCU to PWI, students also experienced difficulties adjusting to the pedagogical style and expectations of the faculty members at the Research I institution in the delivery of graduate instruction:

“Here, it’s a group of instructors, professors that teach the class. So it takes a couple of days to get used to different people training you now. Instead of getting used to one person, you have to get used to a number of people, and they all kind of teach differently. So that took a little bit of time.”

“It took awhile to adjust to teaching styles, different learning styles because at previous institutions we basically had one instructor for the course, and they taught every aspect of that course. Here you have multiple instructors.”

“I guess the teaching style is a little different here. I have never been through a course, under grad or even at the master’s level, where I had three teachers teaching one subject, and that can be a little tricky, because they all have different teaching styles.”

Providing students with clear expectations of the level of work desired and a thorough review of the curriculum should serve as foundational preparation for students, far in advance of their entry into the program. Similar to what was encountered in Carlone and Johnson’s (2007) study, if students encounter challenges in adjusting academically, this could serve as a negative indicator for the instructor, who could unduly disrupt a student’s self-perception as a scientist by expressing some doubts to their ability to successfully complete the program. Students would benefit greatly from adequate time to adjust, which is difficult given the current culture of science driven by a competitive atmosphere.

Status Differentiation
Perhaps the most urgent challenge faced by these students involves the practice of status differentiation. Administrators of intervention programs are often faced with the dilemma of positively promoting such programs and highlighting the successes of their students, while not unduly singling out a particular cohort of students for what can be perceived as special treatment to the rest of the student body. Unfortunately, students in this study experienced a sense of hypervisibility due to the branding of the program in the school. Students expressed a myriad of perceptions based on their daily interactions with others:

“Sometimes when things happen with the minority students, it seems that everyone knows about it. Where with other students, it may not … everyone may not know about it, but with minority students, it seems that everyone knows about something that’s happened … I mean, yeah, for me it is a little bit embarrassing, a little bit embarrassing. You kind of are looking around wondering, ‘Okay who knows?’”

“I just think each individual, each person has to learn how to adjust individually. And I guess everybody is different, some people can and some people can’t. I don’t think that because of my color or because of where I am from I deserve any special privileges. I don’t feel that way. But at the same time, I don’t think I need to be looked at under a microscope either just because I don’t make an A or just because I’m not in class or just because I’m not doing work the way my committee feels like I should be or whatever.”

“All right, well I think there are some things that are being done now on the front end to try and be more supportive of not only students of color but of all graduate students. Because we do all have similar issues. Ours I feel like are exacerbated a little bit more, they are all out in the forefront, and we are put on crazy person watch. ... Whereas some other students are, ‘Well we’ll handle this,’ and they are a little more discreet about it. But I think our business is completely out there. ... And there is [sic] still people after four years that call us by other people’s names, because they don’t … I guess maybe they can’t tell the difference.”

This perception, of feeling as though one or all URM students are under a microscope, led students to believe that there existed a lack of discretion on the part of their preceptors in relationship to their degree progress and to experience a lack of anonymity not experienced by their white peers. Clearly, receiving this additional attention, though well-intended, did not aid in their socialization into the culture of science in the medical school and may have served as a disruption for some in their development of a science identity, if they did not feel as though they were being accepted as a “science person” on the merits of their work.
DISCUSSION

Given its design, a study such as this is not without its limitations. A phenomenological approach was used to interpret the data within the context of the existing literature and in light of some broad, guiding questions. As such, the findings are not intended to be generalizable but rather to shed light on the experiences of a particular set of student experiences during a particular time period. In addition, as is typical in anthropological research, the open coding of data conducted by the author as the sole researcher places limits on the degree to which analytical validity and reliability can be assessed. However, qualitative studies such as this one are usually conducted to provide insight into what is possible in a particular social context (Chase, 2008, p. 79) and perhaps inspire others who may seek to expand the scope of inquiry and quantitatively test the emergent themes found therein. Finally, this study relies upon document analysis, and thus it captures student experiences as they were relayed to another person (the external reviewer) and not directly to the author. This method was chosen based on convenience, given that several of the study participants were either at their home institutions or had completed their doctorates and were subsequently engaged in postdoctoral work elsewhere. Nevertheless, a more explicitly ethnographic approach would have garnered additional insights through participant observation of students as they were engaged in various activities.3

While the interview transcripts contained responses to questions intended for program assessment, the extracted data nevertheless provide an in-depth and valuable account of student experiences in this particular intervention program, in their own words. The student narratives provide an intimate and representative account of their perceptions of themselves, the program, and their experiences therein as participants.

Overall, the findings from this study reveal that students had both positive and negative experiences in the program and that their peers, mentors, and professional development opportunities were critical to their persistence and success in the program. The participants benefited greatly from the opportunity to engage in summer research. They also relied particularly on a strong network of peers, which developed over time with the persistence of the grant programs. However, another key finding was the unique experiences of hypervisibility that the participants encountered as URM students at a PWI.

With respect to the cohort experience over time, Cohorts 1–3 emphasized challenges with transitioning into graduate school, inclusive of the adjustment to a change in environment (social and physical). Cohorts 4–6 experienced a climate wherein critical mass had developed, and thus their experiences highlighted their reliance upon previous cohorts when navigating challenges. Finally, cohort 6 in particular demonstrated increased knowledge of more competitive interventions at other institutions (e.g., availability of more funding). Indeed, one can speculate as to this being the impetus for the two individuals in that cohort eventually matriculating into other institutions for their doctoral work.

Funding from NIH grants remains the most direct pathway to greater access of URM students to the biomedical pipeline. While some of the students in the more recent cohorts have indicated the need for increased funding, the earlier cohorts often cited funding as the principal reason they chose to participate in the program.

The African American female participants in this study experienced struggles in maintaining their academic standing at times and endured financial challenges. Many of these students also had the added responsibility of children, partners, and other dependents, each competing for their attention, compounding their concentrated efforts to succeed at the institution. Yet these same women tapped into resources, most notably themselves as a network, that assisted them in their progression through the program. In the end, three women, out of the total of six students who chose to pursue the PhD at the midwestern urban institution, have obtained their degrees; two have entered into the professorial ranks, and a third is currently a research biologist at a government agency.

In Carlone and Johnson’s (2007, p. 1201) study, students who were women of color looked to their family and community as a source of support and for a sense of recognition as scientists. Many of these students spoke of pressure received from home to “do well.” This desired recognition from individuals outside academia and the realm of science calls into question the traditional forms of recognition, re-casting the practice in terms of “whose recognition matters most.” The present study revealed similar comments from students about having the added support of family at home and knowing that their success in their programs was directly tied to the hopes and the dreams of family, thus serving as a motivational factor when times were difficult.

Yet students also reported feeling out of place, similar to what Gardner (2008) has identified as a feeling of “not fitting the mold.” Attention to the institutional environment into which the students enter, therefore, is an important factor to consider. An optimal institutional environment includes organizational and institutional support for URM students (Oliver et al., 1989). Organizational support ensures a student’s integration into social networks on campus, for example, through student and professional organizations.

Institutional support ensures student interaction with university personnel and entities, particularly when a student is struggling academically.

The institutional environment must also be one that is not only welcoming and nurturing to URM students but also one that is characterized by a strong diversity-valuing climate in which the traditional culture of science is problematized and transformed. An intervention program that seeks to increase the representativeness of its student population must also be attentive to the representativeness of its faculty and staff members (Powell, 2007). If not, it will be accused of not “walking the talk” that it presents. Furthermore, it calls into question the need that URM students often voice about wanting a learning environment in which they see authentic role models who have the potential to motivate them in their pursuit of degrees (Nelson, 2007). It follows that, if there is little to no URM faculty representation, it is easy for URM students to question their identity as scientists and their very presence in a program with the ultimate goal of a PhD, which would gain them entry into the professoriate.
Hypervisibility, as revealed in this study, is a challenge that is difficult to resolve. On the one hand, there is a need to showcase the successes of URM doctoral students, as a means of dispelling myths about the difficulty in mentoring such individuals in relationship to additional time that may be needed and as a means solving the dilemma of underrepresentation in general, so that the culture of science is one that is moving toward more representativeness. On the other hand, there is the problem of singling out a group of individuals for attention that would not otherwise be elicited were it not for their minority status (Beagan, 2003). Majority students in medical schools, defined in terms of the overrepresentation of whites and Asians, do not witness a parallel experience of having the spotlight shown upon them in every department meeting, public lecture, or classroom. Ironically, an added dimension in this dilemma was that of the Bridges students relying upon one another for support; they socialized with each other outside class and attended events collectively. Therefore, their physical presence as distinct group circulating within a predominantly white environment also amplified their hypervisibility.

IMPLICATIONS

The findings of this study, particularly the challenges experienced by the students, suggest a number of improvements that could be initiated for similar programs designed to increase the matriculation, retention, and graduation of URM students in the biomedical sciences. In terms of the opportunity to engage in summer research, the students highlighted the positive contrast between the HBCU and the Research I labs, and that is an important factor to maintain. However, they also seemed to have longed for an added component of summer classes. A program that includes summer research and class work may ease the transition that students may encounter should they choose to matriculate into the partnering institution. In terms of course work, it could also expose them to the difference in pedagogical modes of instruction earlier, which would ease their adjustment. Combining research and course work or having the opportunity to engage in summer research over multiple years supports some findings that suggest a single summer research experience does not create a meaningful experience for students (Fechheimer et al., 2011). The findings of the present study clearly suggest that there needs to be earlier and sustained outreach to students in preparation for the transition from HBCU to PWI.

In terms of mentorship, the type of mentor–protégé contact varied, from group sessions to one-on-one meetings. The student narratives reveal a preference for the latter. Student responses also indicate a greater receptivity toward mentors who demonstrated care in their success, who exercised a degree of nurturing, and/or who could act as an advocate as they progressed through the program. This type of mentorship is also best facilitated through one-on-one contact. Of course, it is also a time-intensive strategy. Programs attentive to these needs may yield greater successes for their URM students.

The students in this study also expressed a desire for authentic role models. These are individuals who are not only URM scholars but who are knowledgeable of their particular experiences (e.g., coming from an MSI). An authentic role model can also be a scholar who can effectively support students through a period of adjustment while inspiring and motivating them to complete a program. Finally, the ability to interact with authentic role models through professional development opportunities (e.g., conference presentations) helps students to envision career pathways they may not have previously considered, particularly in areas of research.

The student responses clearly elucidate the structure that best facilitates their persistence: peer interaction. Programs that build in intentional peer interaction may yield greater results. Intentional peer mentoring could be facilitated through a programmatic matching of previous and current cohorts to ease transition and adjustment. The type of mutual accountability fostered through peer mentorship may go far in solidifying the coconstruction of science identities among students who frankly grow to feel like kin to one another.

Finally, the findings suggest that the highest degree of discretion must be exercised when managing any challenges faced by URM students and that these challenges must be addressed immediately and one-on-one. A step toward developing strong trusting mentoring relationships includes understanding how a student wants to be mentored; specifically, what type of approach is most motivating. The findings suggest that labeling a cohort of students by the grant that supports them may not be an effective means to foster this type of relationship and, indeed, may impede the socialization process.

CONCLUSION

It is important for students to evaluate programs designed to benefit them (Oliver et al., 1989). Through an examination of URM students’ individual and collective experiences as members of a particular program, their voices are effectively heard and contribute to the assessment and evaluation of program effectiveness. With this feedback, program administrators can measure student experiences in relation to the intended goals of the program. Ideally, if the students’ experiences are in line with the intended goals of the program, administrators and coordinators can observe the track of students’ success toward the achievement of the ultimate goal of graduation. On the surface, one can observe this particular Bridges program and gauge its relative success, having obtained the goal of “bridging” 13 out of 18 students into doctoral programs. However, there is much more to be learned, perhaps by listening to the voices that were not included in this study.

The challenges to solving the problem of underrepresentation are multiple and long-standing. While a few students cannot speak for an entire cohort of students, much less capture the entire experience of all URM graduate students enrolled at a particular institution, institutions cannot afford to minimize the importance of their voices either. The experiences revealed in this study relate to the overall experiences documented in previous studies related to URM experiences at PWIs, pertaining to both undergraduate and graduate students. Only by revealing what students experience in their own words can we begin to understand their experiences (Davis et al., 2004). When we understand their experiences, we can affect changes in programs designed to serve them.
ACKNOWLEDGMENTS

This study was made possible in part through funding from the National Institute of General Medical Sciences of the National Institutes of Health (grants R25 GM065792 and R25 GM79657-01). The funding agency made no contribution to the design, data collection, interpretation, or publication of this study. The author acknowledges with appreciation the LSE monitoring editor, Dr. Deborah Allen; the LSE managing editor, Thea Clarke; and the two anonymous reviewers for their careful reading of and helpful suggestions for this article.

REFERENCES


Helping Struggling Students in Introductory Biology: A Peer-Tutoring Approach That Improves Performance, Perception, and Retention

Zachary Batz, Brian J. Olsen, Jonathan Dumont, Farahad Dastoor, and Michelle K. Smith

School of Biology and Ecology and Maine Center for Research in STEM Education, University of Maine at Orono, Orono, ME 04469-5751

Submitted August 7, 2014; Revised January 7, 2015; Accepted January 8, 2015

The high attrition rate among science, technology, engineering, and mathematics (STEM) majors has long been an area of concern for institutions and educational researchers. The transition from introductory to advanced courses has been identified as a particularly "leaky" point along the STEM pipeline, and students who struggle early in an introductory STEM course are predominantly at risk. Peer-tutoring programs offered to all students in a course have been widely found to help STEM students during this critical transition, but hiring a sufficient number of tutors may not be an option for some institutions. As an alternative, this study examines the viability of an optional peer-tutoring program offered to students who are struggling in a large-enrollment, introductory biology course. Struggling students who regularly attended peer tutoring increased exam performance, expert-like perceptions of biology, and course persistence relative to their struggling peers who were not attending the peer-tutoring sessions. The results of this study provide information to instructors who want to design targeted academic assistance for students who are struggling in introductory courses.

INTRODUCTION

The poor retention of undergraduates in science, technology, engineering, and mathematics (STEM) fields has long been an area of concern for U.S. educators and has spurred broad calls to reform undergraduate STEM courses (American Association for the Advancement of Science, 2011; President's Council of Advisors on Science and Technology, 2012). Nationwide, nearly half (48%) of the students who enter a bachelor's program seeking a STEM degree switch into a non-STEM field or leave college altogether (Chen and Soldner, 2013). The majority of students who leave STEM fields decide to do so within the first 2 yr of their program (Watkins and Mazur, 2013). Low grades in introductory STEM courses are often a contributing factor, pushing students out of STEM majors, while higher grades in non-STEM courses pull them toward non-STEM majors (Ost, 2010).

To help retain students in STEM majors, institutions have invested resources trying to improve the performance and retention of students in introductory STEM courses. One commonly used strategy is peer tutoring, in which undergraduates help one another learn within clearly defined tutor and tutee roles (Topping, 2000). A number of national programs have been designed to formalize peer tutoring, including the peer-led team learning (PLTL) and learning assistant (LA)
programs (Gosser et al., 2001; Otero et al., 2010). Both programs train undergraduates who previously succeeded in a course to assist all students currently enrolled in that same course. Peer tutors reduce the student-to-instructor ratio in a course, thereby facilitating additional group work and active-learning opportunities for the students. The introduction of peer tutors into biology, chemistry, engineering, physics, and computer science courses has resulted in increased student grades (Preszler, 2009; Gosser, 2011; Hooker, 2011) and content mastery (Pollock, 2009; University of Colorado Learning Assistant Program, 2011; Chasteen et al., 2012).

While there are many opportunities for peer tutors to assist students in class, they can also help students who are participating in study groups outside class. Typically, the opportunity to participate in these study groups is offered to all students, and the effectiveness of these programs is monitored by comparing course performance metrics for students who participate in the study groups versus those who do not (e.g., Born et al., 2002; Hockings et al., 2008; Stanger-Hall et al., 2010). Because students who join these study groups can have a wide range of abilities, these studies often use demographic information such as incoming grade point average (GPA) or Scholastic Aptitude Test (SAT) score to control for differences in the two populations. After controlling for these differences, several studies have shown that peer-tutoring study groups improve course grades and positively impact persistence (e.g., Born et al., 2002; Hockings et al., 2008, Stanger-Hall et al., 2010). For example, a study offering a PLTL program outside of an introductory chemistry course found that, after controlling for students’ background and other characteristics, students who attended earned a third of a letter grade higher overall in the course than students who declined to attend (Hockings et al., 2008).

The success of peer-tutoring programs in STEM courses suggests that wider implementation could improve student performance and persistence of STEM undergraduates. However, not all institutions have the time and money to provide extra study group programs for every student in a class, especially in high-enrollment introductory STEM courses. Therefore, recent efforts have focused on providing targeted help to students who are struggling early in STEM courses. For example, in a study by Deslauriers et al. (2012), students who performed poorly on the first exam in an introductory physics or introductory oceanography course received an email invitation to meet with the course professor to discuss their current study strategies and reflect on how they might improve them. The struggling students who met with the professor improved their next exam grade nearly 5% more than their struggling peers who had not.

Similarly, a study in an introductory psychology course offered students who failed the first exam an intervention focused on developing metacognitive skills (Lizzio and Wilson, 2013). These struggling students were contacted by a tutor and invited to complete a workbook that was designed to have them reflect on their performance and identify deficiencies in their preparation. Following this self-reflection, students met with the trained tutor for approximately an hour to discuss the contents of their workbook and specific steps they could take to improve. Approximately two-thirds (64%) of students who took part in this intervention passed the course compared with just 27% of the struggling students who declined the intervention opportunity.

In this paper, we build on work demonstrating the benefits of peer tutoring and interventions that target struggling students and ask: Can weekly peer tutoring still be effective in an introductory biology course when it is offered to the lowest-performing students? Specifically, we offered a peer-tutoring program to students who scored 50% or lower on the first introductory biology exam and assessed the impact of this program on exam performance; student perception of biology as a discipline, using the Colorado Learning Attitudes about Science Survey in Bio (CLASS-Bio; Semsar et al., 2011); and persistence. Taken together, this work helps instructors understand how peer tutoring impacts students who are having the most difficult time in a course and how these outcomes can be used to repair STEM’s leaky pipeline.

METHODS
Course Description
This study was conducted in the Introductory Biology (BIO 100) course at the University of Maine (UMaine) during the Fall of 2013. Seven hundred sixty full-time students were enrolled in the course across three sections. Students in all three sections were cotutored by two faculty members (B.J.O. and F.D.), covered the same materials, and took identical exams. Each section met for 50 min, 3 d/wk.

The course is graded out of 1100 points—300 from exams (exams 1–4), 200 from a comprehensive final exam (exam 5), and 100 from weekly online homework assignments and in-class clicker questions. The remaining 500 points come from a required lab component of the course. For this study, exam scores are used as the measure of performance, because they are the primary method of assessment in this course.

Course Demography
The BIO 100 course is the largest course taught at UMaine, and students are primarily first-year, nonminority students from rural counties (Table 1). Most students were enrolled in majors that require additional biology courses. Women outnumbered men almost two to one.

Identification of Struggling Students
In upper-level courses, college GPA is a good predictor of whether a student will struggle (Murtaugh et al., 1999), but this measure was not available, because students often take the BIO 100 course in their first college semester (Table 1). Therefore, in this study, we identified struggling students based on their first exam scores (Jensen and Moore, 2008). The average score on the first exam was 55.4% (SD = 16.7%). For the purpose of this study, students who earned 50% or lower on the first exam were identified as “struggling,” while students who scored above 50% were identified as “non-struggling” (Figure 1).

Shortly after the first exam, struggling students received both an email and a paper invitation to enroll in peer tutoring (Supplemental Figure 1). These invitations expressed concern about the student’s performance on the first exam and contained information about when the sessions would meet, the materials and structure of the sessions, and directions on how to sign up. Approximately one-third of struggling students signed up for the program (Figure 1).
A small number \((n = 6)\) of non-struggling students independently contacted the course professors and asked to join the peer-tutoring program. These students were permitted to join a peer-tutoring session but are excluded from all data analyses.

The undergraduate peer tutors were hired as part of an LA program \((Otero et al., 2010)\). In accordance with the LA program framework, students who previously scored in the top quartile in BIO 100 and expressed interest in a STEM teaching career were hired. The peer tutors were concurrently enrolled in a course covering practical and theoretical aspects of STEM education to help prepare them both for their roles as peer tutors and their potential futures as STEM educators.

**Table 1. Demographic overview of full-time BIO 100 students during Fall 2013**

<table>
<thead>
<tr>
<th>Category</th>
<th>(n)</th>
<th>% of class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>499</td>
<td>65.7</td>
</tr>
<tr>
<td>Men</td>
<td>261</td>
<td>34.3</td>
</tr>
<tr>
<td>Home county</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urban (population &gt; 250,000)</td>
<td>132</td>
<td>18.3</td>
</tr>
<tr>
<td>Rural (population &lt; 250,000)</td>
<td>591</td>
<td>81.7</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonminority</td>
<td>636</td>
<td>91.8</td>
</tr>
<tr>
<td>Minority(^b)</td>
<td>57</td>
<td>8.2</td>
</tr>
<tr>
<td>Additional biology course requirements</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plan requires additional biology</td>
<td>576</td>
<td>75.8</td>
</tr>
<tr>
<td>Plan does not require additional biology</td>
<td>130</td>
<td>17.1</td>
</tr>
<tr>
<td>Undeclared</td>
<td>54</td>
<td>7.1</td>
</tr>
<tr>
<td>Year in school</td>
<td></td>
<td></td>
</tr>
<tr>
<td>First year</td>
<td>576</td>
<td>75.8</td>
</tr>
<tr>
<td>Sophomore</td>
<td>134</td>
<td>17.6</td>
</tr>
<tr>
<td>Junior</td>
<td>40</td>
<td>5.3</td>
</tr>
<tr>
<td>Senior</td>
<td>9</td>
<td>1.2</td>
</tr>
<tr>
<td>Standardized test</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>SAT score(^c)</td>
<td>1070.3</td>
<td>131.0</td>
</tr>
</tbody>
</table>

\(^a\)Total Bio 100 enrollment \(n = 760\) full-time students; class size varies slightly throughout table due to gaps in available demographic information.

\(^b\)Minority students include: Hispanic/Latino, American Indian/Alaska Native, black/African American, and Asian.

\(^c\)And/or SAT score equivalent of ACT score (see Methods).

**Structure of Peer Tutoring**

In this study, peer tutoring included 1) weekly meetings with up to 10 struggling students and the peer tutor and 2) question packets prepared by the researchers and faculty instructors that were given to the students each week. The question packets included five multiple-choice questions selected from exams given in previous years. The selected questions aligned with course learning goals and, when possible, were at the application and analysis levels \((Bloom \textit{et al.}, 1956; Supplemental Figure 2)\). Although the faculty members who teach this course do not post old exams, graded exams are returned each year to students. Consequently, students often have copies of exams from previous years, and the questions have been posted to various websites.

Students worked together in small groups to answer each question and write a justification of why the correct answer was correct along with additional justifications for why each incorrect answer was incorrect. While students worked, peer tutors circulated to monitor their progress and answer questions. Toward the end of the session, the peer tutor would bring the group back together to review their answers and justifications.

To investigate how well the peer tutoring prepared students for exams, three researchers \((F.D., Z.B., \text{and} J.D.)\) independently compared the content of each exam question with the content of the questions asked in the peer-tutoring sessions. Questions that did not have 100% agreement were discussed until there was consensus. In total, 55.6% of the exam questions were similar in content to questions covered in the peer-tutoring sections.

To ensure that peer tutors were prepared for these sessions, they attended weekly meetings with the course instructors to go through the questions before their sessions with students. The peer tutors were expected to complete the question packet on their own before coming to the instructor meeting. During each meeting, the peer tutors collectively reviewed their answers to and justifications for each question.

**Student Data**

This study focuses on comparing the outcomes for struggling students who enrolled in optional peer tutoring with those struggling students who declined to enter peer tutoring. The peer-tutoring program represents one difference between
these two groups, but students may differ in other critical ways. To account for this, we compared these two groups of students using demographic, engagement, performance, and perception metrics. Although it is impossible to rule out significant differences in other, unmeasured aspects, these metrics provide a broad view of the types of students in each group and the factors that impact student performance.

**Demographic Information**

We considered a student’s gender, ethnicity, year in school, and whether his or her home county is urban or rural (U.S. Department of Agriculture, 2003; Table 1). In addition we considered whether a student’s major required additional biology courses in the future and his or her standardized test scores. At UMaine, students can submit scores from the SAT and/or the ACT for admission (Table 2). For comparisons, we converted all ACT scores into SAT scores by first converting ACT scores to z scores using the national ACT statistics (mean = 21.1, SD = 5.3) and then converting these z scores to equivalent SAT scores using the national SAT statistics (mean = 1010, SD = 164.8). The formula for this conversion is:

\[
\text{equivalent SAT score} = \left( \frac{\text{student ACT score} - \text{national ACT mean}}{\text{national ACT SD}} \right) \times \text{national SAT SD} + \text{national SAT mean}
\]

If a student submitted both SAT and ACT scores, his or her ACT score was converted to an SAT score, and the average of that score and the submitted SAT score was used for analysis.

In total, 53 students submitted both scores, and there was no significant difference between the mean true SAT score and the calculated equivalent SAT score (paired t test; mean difference = 12.9, SD = 75.3; t(52) = 1.25, p = 0.22). Students with neither SAT nor ACT scores available were excluded from any analyses in which SAT is used as a covariate.

**Engagement Metrics**

In addition to demographic information, we also collected four metrics from Synapse, a local course management system (CMS). CMS metrics included:

1. Synapse log-ins: The number of times a student logged into the Synapse CMS. This system acts as a hub for all lecture and lab materials and all communications from faculty members, teaching assistants, and peer tutors.
2. Note opening: Skeleton versions of the lecture slideshows were posted online in advance of each lecture. Students were able to access these notes beginning approximately 1 wk before the lecture and then again at any point over the remainder of the semester. Lecture notes were not available via any other method (e.g., they were not handed out in class). We counted each time a student opened the page containing the lecture notes. Notes could be downloaded or printed, so we could not capture the total number of times during the semester that students viewed the notes outside the course management system.
3. Video and study guide viewing: For most lectures, the faculty members posted a condensed video version of the lecture along with an outline of key topics covered that day. The video could be viewed only within Synapse, but the study guide was available for download. These two items were posted on the same page, so we were unable to determine which resource a student accessed. Furthermore, there is no way to determine how long a student was on the page or what proportion, if any, of the video he or she watched. Thus, we counted the number of times a student opened a page containing these two resources as an index of access to both resources. Furthermore, video and study guide views are positively correlated with lecture attendance (Pearson’s r = 0.15, p < 0.001), suggesting that students typically use these resources to supplement lecture rather than to replace it.
4. Grade book access: Student grades for both the lab and lecture components of the course were posted exclusively on Synapse. We counted the number of times each student accessed the grade book. Although it is difficult to know why some students check their grade book more often, this metric may indicate that students are taking an interest in their performance.

In this study, the term “engagement” refers to these four CMS metrics and lecture attendance. Lecture attendance was tracked using an in-class clicker system. A student was marked as attending as long as he or she answered at least one clicker question. Penalties were outlined in the syllabus if students were caught using multiple clickers to cover for friends who were not attending, and informal observation suggests this was not a widespread problem throughout the semester.

All engagement metrics were divided into two time periods 1) preintervention: from the start of the semester to the first exam; and 2) intervention: after the first exam until the end of the semester.

**Course Performance**

In addition to the demographic and engagement data, we also analyzed course performance. For the preintervention period, grades on the first exam were compared between struggling students who accepted and declined peer tutoring to assess whether these groups differed. For reference, the performance of these two groups was compared with non-struggling students who performed > 50% on exam 1.

For the intervention period, scores on exams 2–5 were compared between struggling students who accepted and declined peer tutoring. The exam performance of these two groups is then further compared with non-struggling students to place group performance in the context of the entire class.

**Perceptions**

Student perceptions of biology were compared using results from the CLASS-Bio survey (Semsar et al., 2011). Student
perceptions positively correlate with performance, and the perceptions of biology majors tend to improve over 4 yr of study (Hansen and Birol, 2014). Students took the CLASS-Bio online at the beginning (pretest) and end (posttest) of the semester and responded to 31 items such as, “To understand biology, I sometimes think about my personal experiences and relate them to the topic being analyzed,” on a five-point Likert scale ranging from “strongly disagree” to “strongly agree.” Students received a percent favorable score and were placed along a novice-to-expert continuum based on their alignment with the consensus opinions of biology PhDs.

For the preintervention period, pretest percent favorable scores were compared between the two struggling student groups to determine whether students who accepted peer tutoring held significantly different perceptions of biology than students who declined. As with exam performance, the perceptions of both struggling groups were compared with non-struggling students to provide context.

For tracking the relative changes in student perceptions during the course, percent favorable responses on the CLASS-Bio posttest were compared between struggling students who accepted peer tutoring, struggling students who declined peer tutoring, and non-struggling students. In addition, for students who completed both CLASS-Bio pretest and posttest, paired shifts from pretest to posttest scores were examined for struggling students who accepted and declined peer tutoring. A positive shift indicates that students gained more expert-like perceptions of biology during the semester, while a negative shift indicates they left class with more novice-like views. Previous research has found that students in introductory biology courses typically undergo negative shifts from the beginning to the end of a course (Semsar et al., 2011). While the reasons for these novice-like shifts are still being explored, students in courses that incorporate both active learning and an explicit epistemological curriculum are more likely to show no or expert-like shifts from the beginning to the end of the course (Perkins et al., 2005).

Statistical Models

For determining whether struggling students who accepted peer tutoring were different from students who declined peer tutoring, all demographic (Table 1) and preintervention engagement metrics were input into a single, multivariate, logistic regression. This model answers the question: If all demographic characteristics and measures of engagement are held constant except for factor X, does factor X significantly predict whether a struggling student accepted or declined peer tutoring?

To determine whether peer tutoring improved student performance after exam 1, we modeled a student’s mean score on the four exams taken after the intervention began (exams 2–5) as a function of the number of peer-tutoring sessions each struggling student attended, using a multiple linear regression. Only students who attended at least one peer-tutoring session were included in this model. We input the following covariates to isolate the effect of the peer tutoring: demographic characteristics (Table 1) and the five measures of student engagement collected during the intervention period (four CMS metrics and lecture attendance).

This type of model, which has been used to explore similar education research questions (Hocking et al., 2008; Theobald and Freeman, 2014), identifies which factors are associated with increased exam performance when all other factors are held constant.

Finally, to determine whether students who accept peer tutoring improve their perceptions of biology, we modeled CLASS-Bio posttest scores, which measure percent agreement with experts, with the following covariates: whether a student accepted peer tutoring, demographic characteristics (Table 1), five measures of engagement during intervention, and percent favorable scores on the CLASS-Bio pretest.

We tested the independent variables in the logistic and linear models for high collinearity (Pearson’s |r| > 0.70). Pairs of factors with high correlation are marked in the results. Collinearity may obscure significant input factors in some cases (Dormann et al., 2013), and therefore, in each case, the collinear factors were removed from the model one at a time, and these new models were compared with the old to identify any significant changes in the results.

Persistence

To determine whether persistence in the course was affected by the peer-tutoring intervention, we calculated the percentage of students who dropped the course by determining which students took the mandatory final exam; students who did not take the final exam were marked as dropped. The percentage of students who dropped was compared between struggling students who accepted peer tutoring and struggling students who declined peer tutoring using a Z test for two proportions. Students who dropped out during the course of the semester are excluded from all other analyses to keep groups comparable across the entire study.

Student Surveys

In the final peer-tutoring session, students were given a brief survey to assess their opinions about the peer-tutoring program. These surveys included three Likert-scale items for students to share their opinions on whether the peer-tutoring program helped them develop studying and test-taking skills both in BIO 100 and more generally.

To understand why some students did not participate in peer tutoring, we asked struggling students who declined to participate to fill out a free-response online question that reminded them about the peer tutoring and asked them to describe why they had declined help. Students were told that their answers would help the instructors better design future peer-tutoring programs. Responses were collected from 140 of the 215 struggling students who had declined peer tutoring (65.1%), and the reasons were coded into categories by two independent raters. Some students gave more than one reason in their response and therefore received more than one code. Interrater reliability was found to be acceptable (κ = 0.67, p < 0.001; Landis and Koch, 1977), and the raters discussed discrepancies to reach a consensus coding.

All statistical tests were performed using SPSS Version 21 (SPSS, IBM, Armonk, NY). Approval for this study was obtained from the university’s institutional review board (UMaine IRB 2012-12-14).
RESULTS

Do Struggling Students Who Accept Peer Tutoring Differ before Intervention from Those Who Decline?

BIO 100 students who scored < 50% on the first exam were invited to participate in the peer-tutoring sessions, and approximately one-third of struggling students accepted (Figure 1). To determine whether struggling students who accepted differed from those who declined, we examined the exam 1 performance of different groups before peer tutoring began (Figure 2A). The non-struggling students performed significantly higher when compared with struggling students who accepted and declined peer tutoring. Notably, the exam 1 scores of struggling students who went on to join peer tutoring were not significantly different from the exam 1 scores of students who declined peer tutoring.

In addition to examining exam 1 scores, we also compared CLASS-Bio (Semsar et al., 2011) scores on the pretest given at the start of the semester (Figure 3A). Non-struggling students had significantly higher agreement with experts compared with struggling students who accepted and
declined peer tutoring. Furthermore, struggling students who accepted peer tutoring began the semester having perceptions of biology equivalent to the perceptions of struggling students who declined peer tutoring.

Next, all of our demographic and preintervention engagement variables for the struggling students were included in a logistic regression to determine whether any of our measured characteristics predicted whether or not a struggling student accepted peer tutoring. The resulting model was not statistically significant compared with an intercept-only model ($\chi^2 (11) = 16.77, p = 0.12$), indicating that these input variables do not significantly predict whether or not a student would accept peer tutoring.

Do Struggling Students in the Peer-Tutoring Program Improve Their Course Performance?

Given that students who accepted and declined peer tutoring were similar at the beginning of the course, we next wanted to determine whether the peer-tutoring program helped students improve their performance on subsequent exams. Students who accepted peer tutoring showed improvement in performance relative to struggling students who declined peer tutoring immediately after the intervention, and the differences were consistent throughout the semester (Figure 4). Furthermore, when averaged scores on exams 2–5 were compared, non-struggling students outperformed struggling students; however, the exam performance of struggling students who accepted peer tutoring was significantly better than struggling students who declined peer tutoring (Figure 2B). Moreover, students in the peer-tutoring program with the greatest improvement in their exam performance attended, on average, more peer-tutoring sessions (Supplemental Figure 3).

These results suggest that attending peer tutoring helped struggling students improve their exam performance. However, a number of other factors may explain improved exam performance. Therefore, we ran a multiple regression model to identify the impact of peer-tutoring attendance on exam performance. Therefore, we ran a multiple regression model of factors contributing to mean performance on intervention period exams for students attending at least one peer-tutoring session (Table 3).

After controlling for differences in demographic characteristics and student engagement, the model predicts that struggling students who accepted peer tutoring would gain 0.99 percentage points on their mean 2–5 exam scores for each tutoring session they attended ($p = 0.03$; Table 3). There were 10 available sessions, so struggling students who attended all weekly sessions would be expected to score approximately one letter grade higher on these last four exams on average than struggling students who attended just one peer-tutoring session.

Table 3. Multiple regression model of factors contributing to mean performance on intervention period exams for students attending at least one peer-tutoring session

<table>
<thead>
<tr>
<th>Factor</th>
<th>$\beta \pm SE^a$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peer-tutoring program</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of sessions attended</td>
<td>0.99 ± 0.44</td>
<td>0.03</td>
</tr>
<tr>
<td>Engagement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Video/study guide views</td>
<td>0.15 ± 0.05</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Lectures attended</td>
<td>0.53 ± 0.23</td>
<td>0.02</td>
</tr>
<tr>
<td>Note openings</td>
<td>−0.05 ± 0.06</td>
<td>0.35</td>
</tr>
<tr>
<td>Grade book checks$^b$</td>
<td>0.02 ± 0.04</td>
<td>0.64</td>
</tr>
<tr>
<td>Synapse log-ins$^b$</td>
<td>0.01 ± 0.04</td>
<td>0.79</td>
</tr>
<tr>
<td>Demographics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAT score</td>
<td>0.04 ± 0.01</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Year in school (1 → 2 → 3 → 4)</td>
<td>2.59 ± 2.03</td>
<td>0.20</td>
</tr>
<tr>
<td>Gender (male → female)</td>
<td>2.49 ± 2.83</td>
<td>0.38</td>
</tr>
<tr>
<td>Home county (rural → urban)</td>
<td>2.45 ± 2.82</td>
<td>0.39</td>
</tr>
<tr>
<td>Require more biology? (no → yes)</td>
<td>0.92 ± 2.77</td>
<td>0.74</td>
</tr>
<tr>
<td>Ethnicity (nonminority → minority)</td>
<td>0.39 ± 3.31</td>
<td>0.91</td>
</tr>
</tbody>
</table>

Factors with $p$ values <0.05 are indicated in bold font.

$^a$ Indicates the expected change in the mean exam score (in percentage points) given a unit change in the parameter of interest. For categorical variables, $\beta$ indicates the predicted change in score that results from a categorical change in the direction indicated by the arrows in parentheses (e.g., changing the level of “gender” from “male” to “female” increases the expected mean exam score by 2.49 percentage points).

$^b$These factors are highly correlated ($r = 0.90$), but removing one of them does not change which input variables are significant or the $\beta$ values of significant independent variables. Supplemental Figure 5 shows the results from the multiple regression model when the factor Synapse log-ins is removed.
Average performance on the four intervention period exams was also significantly predicted by video and study guide viewing ($\beta = 0.15$, $p < 0.001$) and lecture attendance ($\beta = 0.53$, $p = 0.02$; Table 3). More frequent viewing of videos and study guides and higher lecture attendance are significantly associated with higher exam scores. Additionally, higher SAT scores are associated with better performance on these exams ($\beta = 0.04$, $p < 0.001$). However, it is notable that, after adjusting for the impact of these factors, peer-tutoring attendance still has a significant impact on exam performance.

Several factors do not significantly impact exam performance after adjusting for other factors ($p \geq 0.20$; Table 3). These factors include: differences in note opening, grade book checks, Synapse log-ins, year in school, gender, urban/rural hometown, future biology requirements for major, and minority status.

### Do Struggling Students in Peer Tutoring Improve Their Perceptions of Biology?

In addition to their poor performance early in the year, struggling students who accepted and declined peer tutoring entered the class holding more novice-like perceptions of biology than their non-struggling peers (Figure 3A). At the end of the semester, struggling students who accepted and declined peer tutoring still held more novice-like perceptions of biology than their non-struggling peers, but struggling students who accepted peer tutoring ended the year with significantly higher CLASS-Bio scores than struggling students who declined (Figure 3B). In addition, struggling students who declined peer tutoring shifted significantly toward more novice-like perceptions of biology from pretest to posttest, while struggling students in peer tutoring maintain their initial perceptions (Figure 3C).

Next, we ran a multiple regression model among the struggling students to determine whether accepting peer tutoring was significantly associated with higher CLASS-Bio posttest scores after controlling for demographic, intervention period engagement, and CLASS-Bio pretest score factors. The majority of students included in this model declined peer tutoring and accordingly attended zero peer-tutoring sessions. To avoid zero inflation in our previously continuous peer-tutoring attendance variable, we defined peer tutoring in this analysis as a binary variable (accept or decline).

The model is significant compared with an intercept-only model ($F(13,176) = 13.86$, $p < 0.001$) and accounts for 46.9% (adjusted $R^2$) of the variance in CLASS-Bio posttest scores (Table 4). After controlling for differences in student engagement, demographic characteristics, and CLASS-Bio pretest score, it was predicted that struggling students who accepted peer tutoring would score 6.05 points on the CLASS-Bio posttest ($p = 0.02$) compared with students who declined peer tutoring.

Also, CLASS-Bio pretest and posttest scores are positively correlated (Table 4). After accounting for all other variables in the model (including peer-tutoring participation) for each point scored on the CLASS-Bio pretest, students scored 0.71 points more favorable on the CLASS-Bio posttest ($p < 0.01$).

In addition, a number of other factors showed interesting but nonsignificant trends ($0.05 < p < 0.2$) on CLASS-Bio posttest scores (Table 4). These trends may be too variable or weak, given the statistical power of our test. For example, female students are predicted to end the semester with more expert-like perceptions ($\beta = 5.05$, $p = 0.08$) compared with similar male students. Other factors approaching significance in the model include number of video/study guide views, the number of lectures attended, SAT scores, and year in school.

### Do Struggling Students in Peer Tutoring Persist at a Higher Rate?

Ultimately, the goal of this tutoring program is to improve the persistence of struggling students in introductory biology. To evaluate this goal, we compared the overall percent of students dropping out of BIO 100 with the dropout rate for each group and found that struggling students who declined peer tutoring had the highest dropout rate of any group (Table 5). Moreover, the dropout rate of struggling students who declined peer tutoring is significantly higher than struggling students who accepted peer tutoring ($Z$ test for two proportions, $Z = 3.25$, $p < 0.001$).

### What Do Struggling Students Think of Peer Tutoring?

During the final peer-tutoring session, a brief survey was given to assess student opinions of the peer-tutoring program.
(Table 6). Students who participated in the peer-tutoring programs overwhelmingly said the peer tutoring was helpful for both the BIO 100 course and their other courses.

**DISCUSSION**

This study used peer tutors to provide targeted help to students who did poorly on the first exam in a large-enrollment, introductory biology course. Students who struggle on the first exam in introductory biology typically fail to improve significantly over the remainder of the semester (Jensen and Moore, 2008). While previous research has described the improvements in student performance, perceptions, and attitudes as a result of offering peer tutoring (Plant et al., 2002; Hockings et al., 2008; Stanger-Hall et al., 2010), we report here how those benefits can be replicated within a subset of struggling students who are participating in ungraded, extracurricular sessions. Notably in this study, struggling students who regularly attended peer-tutoring sessions improved their exam performance (Figures 2 and 4 and Table 3) and persisted at a higher rate (Table 5). Students who participated in peer tutoring also finished the course with more expert-like perceptions of biology (Figure 3B and Table 4) compared with struggling students who declined peer-tutoring sessions.

Because institutions may not have the resources to hire and train enough peer tutors to support class-wide implementation of peer tutoring, it is important to know that peer tutoring offered to struggling students is effective. Furthermore, the improved performance of struggling students who accepted peer tutoring is not the result of a biased self-selection in the measured characteristics (Figures 2A and 3A). It is possible that our metrics do not document some important variation between our groups. However, it is worth noting that, before the widespread adoption of online CMSs like Synapse, the adjustments for student engagement presented here would not have been possible. Given continued advances in classroom technology, future studies may capture additional variations to better characterize the impact of targeted intervention strategies.

**Why Is Peer Tutoring Effective for Struggling Students?**

In this study, peer tutoring refers to a system of student support, including weekly meetings with up to 10 students, a trained peer tutor, and prepared question packets given to the students each week. One plausible explanation for the benefits of this program is that the struggling students who accepted peer tutoring improved their exam performance simply because they were spending an extra hour per week studying. However, previous work has found a negligible relationship between the amount of time spent studying and academic performance (Nonis and Hudson, 2010; Masui et al., 2012), suggesting that the quality of studying techniques is also important.

High-quality study techniques can be described as an extension of the deliberate practice model of expert development (Plant et al., 2005). The deliberate practice framework proposes that developing expertise in a domain requires:

1. Regularly scheduled training in that domain;
2. Training tasks designed by an expert teacher;
3. Ongoing feedback from an expert teacher during training tasks; and
4. Intense concentration during training with minimal distractions (Ericsson, 2006; Ericsson et al., 1993).

The peer-tutoring sessions described in this study were explicitly designed to follow this framework. Struggling students met the first condition by signing up for weekly tutoring sessions. The second condition was met, because the materials provided in these sessions were designed by the course instructors and research investigators to align with course goals. The third condition was also met, because students received feedback and clarification from their trained peer tutors as they worked. Finally, students met the last condition by meeting with peer tutors who monitored the group and encouraged their students to stay on task throughout the session.
that can help students who feel isolated. Isolation from the educational community can lead to disengagement and low persistence among students in large introductory courses (Tinto, 1993; Evans et al., 2001). Learning communities have been found to combat isolation and improve attitudes, learning outcomes, and motivation for students in STEM fields (Freeman et al., 2008).

Peer tutoring is also unique in that it benefits the students acting as tutors. Outcomes for peer tutors were not specifically investigated in this paper, but previous research has consistently shown that peer tutors improve their content mastery as well as their perceptions of the respective field (Otero et al., 2010; University of Colorado Learning Assistant Program, 2011). In this way, a targeted peer-tutoring program maximizes the benefits of limited resources by significantly improving the performance of struggling students in an introductory course, by increasing retention of STEM majors, and by helping peer tutors develop teaching skills and further mastery in a STEM field.

**Reaching Students Who Declined Peer Tutoring**

Peer tutoring can improve the performance and persistence of struggling introductory biology students, but this optional program can only help those who choose to attend. This self-selection process means that students who enrolled in the program could have potentially been quite different from struggling students who did not. However, analysis of the performance, perceptions, demographic characteristics, and engagement of the students in these two groups

---

**Figure 5. Reasons struggling students give for declining peer tutoring.** Some students gave more than one reason in their response and therefore received more than one code.
showed they were similar to the students who declined the assistance (Figures 2A and 3A). These results suggest that, while the background, engagement, and early performance of STEM students are important (Tinto, 1993; Seymour, 2000; Ost, 2010; Watkins and Mazur, 2013), the measures presented here are not sufficient for determining which students are likely to accept help.

Despite the benefits of the peer-tutoring program, not all struggling students participate. For example, in this study, 35.2% of students accepted the offered peer tutoring after struggling on the first exam (Figure 1). Previous studies offering ungraded, extracurricular help to struggling students in introductory psychology courses (Lizzio and Wilson, 2013), struggling students in introductory physics and oceanography courses (Deslauriers et al., 2012), and all students in an introductory chemistry course (Hockings et al., 2008) had acceptance rates of 40.5, 39.1, and 40.0%, respectively. The similarity of these acceptance rates suggests that student interest in this type of assistance may be somewhat constant over a range of classroom settings.

The participation rates for struggling students leads to this question: Why are some students not interested in taking advantage of available extra help? For this peer-tutoring program, many students reported they were unable to make time for a weekly meeting, did not see that they needed help, or chose other available help in lieu of the peer-tutoring program (Figure 5). Accordingly, we suggest that future recruitment efforts should highlight for students the differential exam performance between struggling students in and out of peer tutoring. Furthermore, students frequently reported scheduling conflicts during available tutoring times; future registration systems should include a wait-list component so program directors are aware of additional students wanting to register for a popular time slot and can shift peer tutors from less popular times to meet demand.

CONCLUSION

This study shows that some students who do poorly on the first exam in a course will participate in a targeted tutoring program, even though the program is ungraded and requires additional time investment. Students who accepted peer-tutoring sessions achieved improved exam performance and ended the year with more expert-like perceptions of biology relative to their struggling peers who declined help. Ultimately, these improvements led to increased persistence in the class and suggest that targeted peer tutoring can be an effective tool for reducing the loss of at-risk students during the critical, early undergraduate portion of the STEM pipeline.

ACKNOWLEDGMENTS

The authors appreciate the participation by the BIO 100 students. We thank Mary Tyler, Karen Pelletreau, and Natasha Speer for helpful feedback on the manuscript. We also thank Erin Vinson for assistance coding student surveys. Thank you to Ryan Cowan, Ron Kozlowski, and the rest of the Synapse staff for assembling data for this analysis. This work is supported by National Science Foundation grant 0962805.

REFERENCES

American Association for the Advancement of Science (2011). Vision and Change in Undergraduate Biology Education: A Call to Action, Washington, DC.


Use of Feedback-Oriented Online Exercises to Help Physiology Students Construct Well-Organized Answers to Short-Answer Questions

Jacqueline Carnegie

Department of Cellular & Molecular Medicine, University of Ottawa, Ottawa, ON K1H 8M5, Canada

Submitted August 21, 2014; Revised March 25, 2015; Accepted April 7, 2015

Monitoring Editor: Michèle Shuster

INTRODUCTION

Many science disciplines require both a critical arsenal of factual knowledge and an in-depth conceptual understanding so that students can apply their knowledge and understanding to novel problems and contexts (Crowe et al., 2008; Mynlieff et al., 2014). It is the ability of adult learners to use what is learned during lectures to critically evaluate a novel situation and problem solve that is widely recognized as an important goal of university-level teaching of physiology as well as other basic sciences such as biology and chemistry (Pratt, 1993; Zoller, 1993; Tanner and Allen, 2005; Levesque, 2011). Bloom’s revised taxonomy uses six key verbs (remember, understand, apply, analyze, evaluate, and create) to categorize levels of learning; each verb denotes a successively higher level of cognitive ability (Anderson et al., 2001). During the initial years of university, large class sizes and the introductory nature of course content necessitates that assessment of student learning largely take the form of multiple-choice questions (MCQs) that test primarily the lower-order cognitive skills (LOCS) of remembering and understanding. However, as students move through their years of postsecondary education into upper-division courses like physiology, they should be increasingly challenged to develop and display not only LOCS but also the higher-order cognitive skills (HOCS)—analyze, evaluate, and create—as applied to novel or unfamiliar new problems and apply what they have previously learned toward the development of solutions (Zoller, 1993; Michael, 2006; Mayer, 2008; Orr and Foster, 2013; Mynlieff et al., 2014). These abilities fall under the domain of scientific literacy, a competence that should be an important focus of university undergraduate education (Norris and Phillips, 2003; Libarkin and Ording, 2012).
Students progressing through basic science programs often find it difficult to make the transition from answering MCQs that assess primarily remembering, understanding of individual pieces of information, essentially a passive activity, to developing well-worded answers that demonstrate their ability to both apply basic science principles and link them in a logical manner, which is an active process (Zoller, 1993; Bailin, 2002; Crowe et al., 2008). Higher-order learning involves the reorganization of new knowledge and understanding as it is transferred from working memory into long-term memory in a way that forges links with that which is already known and understood so that each student’s long-term memory can evolve as it integrates new components associated with a more complex understanding of scientific processes (Kirschner, 2002; Kirschner et al., 2006). Referred to as “generative processing” (Wittrock, 1992; Mayer, 2010), this type of higher-order learning can be facilitated (Chi et al., 1994) and assessed by asking students to construct concise, well-written answers to short-answer questions (SAQs). For students with limited experience in this type of evaluation, this can be challenging. Indeed, I have noted the following when marking answers to SAQs provided by students during previous years of summative physiology examinations: inaccurate information, the provision of information that, while correct, is not relevant, the omission of key points that would demonstrate an in-depth understanding of relevant concepts, and, finally, an inability to organize the answer logically. I concluded that students needed to practice developing well-organized and accurate answers to SAQs and that, as instructor, I needed to provide that practice before students encounter similar styled questions on summative examinations (Gagné et al., 1992; Levesque, 2011). Unfortunately, the large sizes of many undergraduate classes make it difficult to offer regular practice in developing explanatory answers and solving problems and, more importantly, to provide students with timely, answer-specific feedback on the written work they have prepared.

It is also important to match course-associated learning activities with summative assessment criteria so that, through practice, students might gain or improve their ability to use HOCs before their work is evaluated via high-stakes summative examinations (Sundberg, 2002; Tanner and Allen, 2004; Bissell and Lemons, 2006; Crowe et al., 2008). Indeed, studies have shown that assessment methodology influences the strategies used by students when preparing to write examinations (Scoular and Prosser, 1994). For example, if it is known that an examination will be exclusively MCQ-based, students will often prepare for it by using surface strategies such as the compilation and memorization of lists of factual information, because they know that assessment methods will be used that address primarily LOCS (Scoular, 1998). However, if students know that the summative assessment has been designed to test HOCs such as analysis, evaluation, and the provision of detailed explanations, then students will use learning strategies that involve the use of HOCs through in-class and out-of-class activities that support these approaches (Scoular, 1998; Crowe et al., 2008).

The purpose of this project was to develop an online tool that would allow students enrolled in a third-year level physiology course to practice building, step-by-step, well-organized and comprehensive answers to questions housed within clinical situations or specific physiological scenarios. The development of these interactive exercises was guided by Robert Gagné’s nine events of instruction (Figure 1) so that the ability of this tool to capture student attention and promote interactive, effective learning could be maximized (Gagné et al., 1992). Students will not use an online tool simply because it exists; rather, they must see value in it, including how it links to course objectives and has the potential to improve their summative assessment outcomes (Saunders and Gale, 2012). While it has been more than 20 yr since its original publication, Gagné’s list of instructional goals continues to influence instructional design, because it addresses basic principles, such as the need to motivate and engage learners as well as to guide them toward integrating new knowledge with that which is already known so as to enhance understanding, retention, and transfer (Kinzie, 2005; O’Byrne et al., 2008; Carnegie, 2013). The exercises developed for my study asked students to demonstrate their understanding of four important physiological concepts: 1) the regulation of metabolic rate by thyroid hormone, 2) the control of blood sugar by insulin, 3) the influence of altitude on the regulation of erythropoiesis, and 4) the intrinsic regulation of cardiac stroke volume. While the questions used for these online exercises were not identical to questions students subsequently encountered on midterm examinations, they were representative of the level of cognitive ability the students would be expected to exhibit in the short essay component of summative evaluations.
of each of the four online assignments, students were provided with guidance on answer construction and instantaneous feedback regarding the appropriateness of the items they had selected to include in their answers. At the end of the course, students were asked to complete an anonymous survey detailing their perceptions of the effectiveness of the online tool. In addition, student outcomes on SAQ sections of summative exams were compared between populations that had access to the online tool and those who did not and between subgroups of students who showed various levels of participation in the online SAQ assignments.

METHODS

Exercise Creation

PHS3240 is a full-year, third year–level course in mammalian physiology that is team taught by a number of faculty members within the Department of Cellular and Molecular Medicine at the University of Ottawa. Systems within this course that are taught exclusively by the author include the endocrine and cardiovascular systems. Students learn about the endocrine system during the first part of the course and are assessed on that knowledge in exam 1. Similarly, the cardiovascular system is covered during the next section of the course, with summative assessment of that content during exam 2. Examination scores for this course are typically derived in equal parts from MCQs and SAQs, with each SAQ having a value of five to seven points.

Four interactive exercises were developed for students taking PHS3240, using the action maze software Quandary (Arneil and Holmes, 2009). This freeware can be downloaded for use without charge at www.halfbakedsoftware.com/quandary_download.php. Advantages of this software include the ability to send students back to retry a question if an incorrect answer is selected and to provide feedback that individually targets every answer choice, be it correct or incorrect. Furthermore, the scoring function of Quandary allows points to be awarded or deducted, as appropriate, and students to see their scores as they proceed through the exercise. Initial exercise planning involved the use of tables to organize the questions, the answer choices, the answer-specific explanatory feedback, and the marking scheme. Interactive exercises were then created by loading that information into Quandary, defining the navigation to be followed by the student, and finalizing the scoring. The first two exercises addressed the endocrine system and were completed by students in preparation for exam 1. The second pair of exercises targeted the cardiovascular system and was completed as students prepared for exam 2.

Each exercise opened with the presentation of a physiological scenario (e.g., training for a cycling competition at high altitude) or a clinical vignette (e.g., a patient displaying symptoms of Graves disease) accompanied by one or more essay-style questions for which students would normally be expected to compose a well-organized and concise written answer (Table 1A). However, rather than being asked to provide a written answer, a student was instead taken through a series of six to eight MCQs in which he or she developed an outline of his or her written answer by clicking on the “submit” button to select items that should be included in that answer and, by virtue of the choices he or she made, specifying the order in which that information should be presented. Distractors that were provided in the MCQs included incorrect information, information that was correct but did not apply to the question being answered, and information that was correct but should not be presented until later in the answer after the groundwork for that particular piece of information had been laid (Table 1A).

Feedback was provided for each answer choice, either correct or incorrect. If the answer was correct, the feedback provided a short corroborative explanation, sometimes accompanied by additional pieces of information to further consolidate the concept (Table 1B). If one of the distractors was chosen, the feedback explained why that answer was not the best choice, sometimes provided a hint, and always directed the student back to try again by having him or her click on the “Submit” button associated with that feedback (Table 1C). As an additional effort to reinforce guidance of student thinking (Gagné event 5), the last page of each exercise provided a summary of the correct answers and/or feedback in an effort to show students a worked example of an orderly outline of the elements that should be included in a complete answer to that SAQ (Table 2).

In addition to the inclusion of the summary page, care was taken throughout the process of exercise construction to apply Gagné’s nine events of instruction (Figure 1), as detailed in the Reasoning annotations applied to the working document for exercise creation shown in Table 3. The example used is the endocrine SAQ exercise pertaining to the thyroid gland, and shaded Reasoning boxes shown throughout the table identify examples of specific application of Gagné’s instructional design principles. For example, the patient presenting with Graves disease was used as stimulus material at the beginning of the exercise to grab student attention (events 1 and 4) and provide an opportunity for application of knowledge with the goal of enhancing retention and transfer (event 9). Furthermore, the broad questions that appeared at the end of each case description informed students of the objectives to be addressed within that exercise (event 2). Finally, as detailed throughout Table 3, each of the individual MCQs with their answer choices and associated feedback repeatedly addressed Gagné’s remaining events by stimulating recall, providing practice and learner guidance, and assessment and feedback for every answer selected. Finally, the series of MCQs were constructed in such a way that they guided student thinking as to the logical order in which correct items should appear in their answers (event 5).

Student Populations, Participation, and Feedback

Student enrollment in PHS3240 numbered 144 and 120, respectively, at the beginning of the 2010 and 2011 academic years. There was some attrition over the duration of the course, resulting in 139 and 108 students, respectively, writing exam 2 in the Winter terms of 2011 and 2012. The SAQ exercises were provided online to students using the assignment function of the Blackboard Vista (Blackboard, Washington, DC) course website. Students had 1 wk to complete each exercise and to submit their scores online using the course website assignment drop box. Each correct answer was worth one point, and this resulted in each assignment having a total possible score of six to eight points, depending on the number of questions in that assignment.
Table 1. Online appearance of SAQ interactive exercise pertaining to the regulation of hematopoiesis

A. The SAQ

G.C. is planning to compete in a 100-km cycling race in the mountains in the middle of July. Because he normally lives at sea level, he has decided that it would be a good idea to travel to the race site 6 wk ahead of time to complete his final training. Use your knowledge of aspects of the regulation of hematopoiesis to explain why this training strategy makes sense. If you did a reticulocyte count on G.C. a week or so after he began training in the mountains, what would you find? Related to this, should G.C. have any concerns about a cycling race of long duration at high altitude in midsummer?

2. You have stated that oxygen partial pressure will be reduced at the higher altitude. How will you relate this to the regulation of erythropoiesis?

- a. As G.C. trains at the higher altitude, his muscles will sense reduced oxygen delivery and become larger and stronger.
- b. As G.C. trains at the higher altitude, his kidneys will sense reduced oxygen delivery and stimulate erythropoiesis.
- c. As G.C. trains at the higher altitude, his heart will sense a reduced hematocrit and increase its release of the hormone erythropoietin.
- d. As G.C. trains at the higher altitude, the new red blood cells developing in his bone marrow will switch to producing a type of hemoglobin that can carry a higher number of oxygen molecules.
- e. As G.C. trains at higher altitude, his lungs will increase in size to improve ventilation under conditions of reduced oxygen availability.

B. Feedback for selection of the correct answer to MCQ 2

- b. As G.C. trains at the higher altitude, his kidneys will sense reduced oxygen delivery and stimulate erythropoiesis.

Exactly right. The kidneys are the site of production of the hormone erythropoietin (EPO). When they sense reduced oxygen delivery, they will increase their rate of release of EPO and the higher levels of EPO will stimulate increased rate of production of red blood cells in the bone marrow (erythropoiesis). Please continue to question 3.

C. Feedback for selection of an incorrect answer to MCQ 2

- a. As G.C. trains at the higher altitude, his muscles will sense reduced oxygen delivery and become larger and stronger.

No. There will be reduced oxygen delivery to his muscles, but that will not stimulate them to become larger. They will compensate in other ways, and you may want to think about that. But this statement does not belong in your answer. It is not correct and it is not related to the regulation of erythropoiesis. Please go back and try again.

Analysis of Student Outcomes

The possible influence of online SAQ assignments on student outcomes was assessed by comparing student SAQ scores achieved during the 2 yr before implementation of the online exercises with those earned during the subsequent 2 yr, when students had access to these assignments. A marking rubric was used for each summative examination SAQ, and these rubrics, like the summary pages viewed by students upon the completion of each online exercise, consisted of short lists of required answer components. The same rubric was used each time a particular SAQ was included in a summative examination. Within each rubric, points were awarded not only for including the required answer elements but also for defining and/or explaining them completely and for applying them correctly to the solution of the SAQ. The statistical comparison of student outcomes was conducted using the independent-samples t test accompanied by Levene’s test for equality of variances. Possible
influences of the level of participation in the SAQ assignments on student MCQ and SAQ outcomes were evaluated using analysis of variance (ANOVA) followed by the Tukey range test to identify statistically significant differences between MCQ mean scores and SAQ mean scores for the three levels of participation. For all statistical evaluations, differences were considered significant at \( p < 0.05 \).

RESULTS

Level of Participation

During the 2010–2011 academic year, all of the 144 students enrolled in PHS3240 wrote exam 1. A small number of students dropped the course, resulting in 139 of them (96.5%) remaining to write the subsequent exam (exam 2) during the Winter term. Similarly, 120 (100% of students) completed exam 1 during the 2011–2012 academic year with enrollment dropping to 110 (91.7%) before exam 2. In general, these students showed a high level of participation in the online assignments. Close to 90% (87.5% and 89.2%, respectively, for 2010 and 2011) of students completed the first pair of exercises while studying the endocrine system. However, while 95% of students also completed both of the cardiovascular system–associated exercises during the 2011 Winter term, the level of participation dropped to 82.7% for students given the same assignments during the subsequent academic year. A very small number of students (1.46% and 3.64%, respectively, in each of the two successive academic years) chose not to do any of the exercises at all. In contrast, 84.2% of the 139 students who completed both exams during 2010–2011 participated fully in this aspect of the course, and 74.5% of the 108 students who completed PHS3240 during the second academic year finished all four exercises.

Student Feedback

Survey data were obtained from 71.9% of students who completed PHS3240 during the 2010–2011 academic year. The survey respondents reported that they had completed either three (7%) or all four (93%) of the exercises. The majority of respondents (86%) were in their third or fourth year of undergraduate study, and while many (78%) reported prior experience with SAQs on 50% or more of their summative exams, for the rest of the students (22%), this was their first attempt at composing written work under conditions of exam-related stress.

Figure 2 summarizes student feedback regarding the SAQ exercise content and design. The majority of respondents felt that the online exercises did help them improve their approach to SAQs. Interestingly, there was a rather uniform distribution of students with regard to their perception of the level of difficulty of the exercises. While 58% of students showed partial to complete agreement with the statement that the exercises were sufficiently challenging, the remainder (42%) felt that the exercises could have been constructed to be even more demanding. Students (87%) were especially appreciative of the feedback that was provided for both correct and incorrect answer choices. Finally, the majority of students (78%) indicated that they would have welcomed the opportunity to complete even more of these exercises before writing summative examinations.

Student Outcomes

A comparison of student outcomes on SAQs pertaining to the endocrine and cardiovascular systems (exams 1 and 2, respectively) between students writing exams during the 2 yr before introduction of the online SAQ exercises and during the subsequent 2-yr period when they did have access to
Table 3. Sample working document used for the creation of the SAQ exercises

| **The question:** A 36-yr-old women arrives for her annual checkup with her family doctor. While it is a chilly November day, her doctor is surprised to note that she arrives wearing only a thin sweater, rather than a warm winter coat. She immediately remarks that she feels quite warm and removes the sweater. During the physical exam, the doctor notes an increased heart rate and nervous agitation and makes a note to evaluate her for Graves disease. What is Graves disease and how does this disease account for the observations made by the doctor? If this woman does indeed have Graves disease, what are her treatment options? (8 points) |
| **Reasoning:** Use of a clinical scenario gains attention and presents stimulus material (events 1 and 4). Description of the patient’s symptoms and the identification of her problem as Graves disease stimulates recall of prior learning (event 3) and encourages knowledge application within a clinical context (event 9). Presentation of broad questions at the end of the paragraph informs the learner of the objectives to be addressed in this exercise (event 2). |

1. What is your opening statement as you begin to answer this question?

<table>
<thead>
<tr>
<th><strong>Answer choices</strong></th>
<th><strong>Feedback</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>a. This woman needs to be treated for hyperthyroidism.</td>
<td>This is not the best way to start. You should begin by defining Graves disease and describing the physiological basis of its effects on endocrine function. Please go back and try again.</td>
</tr>
<tr>
<td>b. Graves disease is an autoimmune disease in which the thyroid gland is attacked and destroyed.</td>
<td>This is partly the right way to start, in that you are attempting to define Graves disease, but your information is not completely correct. While Graves disease is an autoimmune disease, it does not result in destruction of the thyroid gland. Please go back and try again.</td>
</tr>
<tr>
<td>c. Graves disease is characterized by high fever and increased release of stress hormones.</td>
<td>While you have the right idea, in that you should begin your answer by defining Graves disease, your information regarding this endocrine disease is not correct. Graves disease is not an infection (so there is no fever) and it does not target the adrenal glands (so it does not directly cause increased release of stress hormones). Please go back and try again.</td>
</tr>
<tr>
<td>d. Graves disease is associated with the production of antibodies that mimic TSH.</td>
<td>Yes. This is a good start. You should also use the word “autoimmune” when describing what kind of disease it is and you should spell out what TSH is. Please continue to question 2.</td>
</tr>
<tr>
<td>e. Graves disease is a disease that disrupts the ability of the hypothalamus to regulate body temperature.</td>
<td>It is good that you are starting by trying to define Graves disease, but the information you are providing is not correct. Body temperature is increased in a patient with Graves disease, but the effect is not at the level of the hypothalamic thermostat. It targets an endocrine gland. Please go back and try again.</td>
</tr>
</tbody>
</table>

**Reasoning:** The question itself provides guidance (event 5) by asking the student to think about the order in which information should be presented. Distractors provide students with information that they should recognize as incorrect (b, c, and e) and address a tendency of some students to struggle with answer organization (a), in this way addressing events 3, 5, and 6. Feedback for the correct answer (d) provides additional guidance (event 5) by suggesting inclusion of the term “autoimmune” in the answer and advising that acronyms must be spelled out in full (a common student error).

2. What are the important pieces of information regarding the thyroid gland that you should include next in your answer? (Choose 2.)

<table>
<thead>
<tr>
<th><strong>Answer choices</strong></th>
<th><strong>Feedback</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Thyroid hormone increases oxygen consumption by virtually all cells of the body.</td>
<td>Well done. This is one of the important facts about the thyroid gland that you should include as the next part of your answer. By stimulating increased oxygen consumption by tissue cells in general, thyroid hormone increases heat production by cells, helping us to maintain body temperature. Please go back and continue choosing correct answers to this question until you have two of them.</td>
</tr>
<tr>
<td>b. Thyroid hormone resets the hypothalamic thermostat to a higher value.</td>
<td>No, this is not one of the functions of thyroid hormone. While the hypothalamus does contain a thermostat, and our body temperature will rise if this thermostat is set to a higher value (e.g., if you have a fever), this is not what is going on in Graves disease. Please go back and try again.</td>
</tr>
<tr>
<td>c. Thyroid hormone inhibits the body’s heat-loss mechanisms.</td>
<td>No, this is not a function of thyroid hormone. While body temperature would rise if we were unable to unload extra heat via peripheral vasodilation and sweating, thyroid hormone does not inhibit these normal heat-loss mechanisms of the body. Please go back and try again.</td>
</tr>
<tr>
<td>d. Thyroid hormone is released by the anterior pituitary.</td>
<td>No. Thyroid hormone is released by the thyroid gland. The hormone released by the anterior pituitary that regulates thyroid gland function is TSH (thyroid-stimulating hormone). Please go back and try again.</td>
</tr>
<tr>
<td>e. Thyroid hormone increases the activity of the Na+/K+ ATPase in many cells of the body.</td>
<td>Yes. This is an important function of thyroid hormone. It stimulates the Na+/K+ ATPase to keep putting Na+ and K+ back in the right place so that cells continue to be responsive. Running the Na+/K+ ATPase requires ATP, so this ties in well with the other important function of thyroid hormone presented in this question: it increases the production of ATP via aerobic metabolism. Please go back and continue choosing answers until you have two that are correct; then proceed to question 3.</td>
</tr>
</tbody>
</table>

**Reasoning:** The goal of this question is to demonstrate the importance of including some basic science at the beginning of the answer to show student understanding of thyroid gland function and its relevance to this clinical problem (events 3 and 5). The three distractors provide incorrect information with the feedback stimulating recall of prior learning (event 3).
In Graves disease, increased nervous agitation is a consequence of the hypersecretion of thyroid hormone by the thyroid gland. While it is true that iodine is required for the synthesis of thyroid hormone, this is "off-topic" information that should not be included in your answer. It does not relate to Graves disease and it does not relate to the general concept of regulation of thyroid hormone synthesis. Please go back and try again.

### Answer choices

<table>
<thead>
<tr>
<th>Answer choices</th>
<th>Feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Thyroid hormone secretion is regulated by the body’s nutritional status.</td>
<td>No. Thyroid hormone influences metabolic rate and so hyper- or hyposecretion of this hormone can have influences on body weight. But the rate of thyroid hormone secretion is not regulated by the body’s nutritional status. Please go back and try again.</td>
</tr>
<tr>
<td>b. The production of thyroid hormone requires iodine.</td>
<td>While it is true that iodine is required for the synthesis of thyroid hormone, this is “off-topic” information that should not be included in your answer. It does not relate to Graves disease and it does not relate to the general concept of regulation of thyroid hormone synthesis. Please go back and try again.</td>
</tr>
<tr>
<td>c. The production and secretion of thyroid hormone is regulated by the hypothalamic-pituitary axis.</td>
<td>Yes. You want to talk about TRH (thyrotropin-releasing hormone) and TSH and how these products of the hypothalamus and anterior pituitary, respectively, regulate the synthesis and secretion of thyroid hormone. You will also want to mention cell surface receptors for TSH on thyroid follicular cells. That will set you up for talking about immune system dysfunction and the production of antibodies that mimic TSH. Please continue to question 4.</td>
</tr>
<tr>
<td>d. The thyroid gland stores several months’ worth of thyroid hormone.</td>
<td>While this is a correct statement, it is extraneous information that does not relate to your answer to a question concerning Graves disease. You want to talk about the hormonal regulation of thyroid hormone synthesis and secretion, not the storage of preformed hormone precursor. Please go back and try again.</td>
</tr>
<tr>
<td>e. Thyroid hormone is secreted only during the cold months of the year.</td>
<td>No. First of all, this statement is not correct. While we do secrete higher levels of thyroid hormone during the colder months, particularly if we are often outside, we also produce and secrete thyroid hormone year-round. Thyroid hormone is needed continually to regulate metabolic rate and to stimulate the functioning of Na⁺/K⁺ ATPase so that our cells are able to maintain a normal distribution of these cations across the cell membrane and so that our excitable cells remain responsive. Please go back and try again.</td>
</tr>
</tbody>
</table>

### Reasoning:

Students need to lay the groundwork for subsequent explanation of uncontrolled stimulation of the thyroid gland in Graves disease by first briefly describing the normal regulation of thyroid hormone secretion (events 3 and 5). Two distractors (b and d) provide guidance to students about the importance of being concise and including only relevant information (events 5 and 7) by providing information that is correct but does not pertain to the question being asked.

### 4. Where does the concept of negative feedback fit into your answer to this question?

<table>
<thead>
<tr>
<th>Answer choices</th>
<th>Feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. In Graves disease, increased body temperature exerts negative feedback to decrease the secretion of thyroid hormone by the thyroid gland.</td>
<td>No. In Graves disease, there is actually hypersecretion, not hyposecretion, of thyroid hormone, and it is the hypersecretion of thyroid hormone that is continually increasing metabolic rate and, therefore, driving body temperature up. Please go back and try again.</td>
</tr>
<tr>
<td>b. In Graves disease, there is a loss of negative feedback that normally controls thyroid gland release of thyroid hormone.</td>
<td>Yes. You should talk about the antibodies that are produced by the immune system being able to substitute for TSH at the TSH receptor on thyroid follicle cells. And then talk about how antibody production would not be adversely affected by rising levels of thyroid hormone, whereas TSH production would be. The absence of negative feedback on antibody production would allow the level of stimulation of thyroid follicle cells to continually increase. Well done. Please continue to question 5.</td>
</tr>
<tr>
<td>c. In Graves disease, there is a loss of negative feedback from the heart on the secretion of thyroid hormone by the thyroid gland.</td>
<td>No. The heart does not produce any hormones that regulate the release of thyroid hormone. While excess thyroid hormone can result in increased heart rate, heart rate does not regulate thyroid hormone secretion. Please go back and try again.</td>
</tr>
<tr>
<td>d. In Graves disease, antibodies are produced that bind to thyroid hormone, preventing it from exerting negative feedback to regulate its secretion.</td>
<td>No. Production of antibodies occurs in Graves disease, but these are antibodies that mimic TSH and stimulate thyroid hormone production and secretion. These antibodies do not interact directly with thyroid hormone. Please go back and try again.</td>
</tr>
<tr>
<td>e. In Graves disease, nervous agitation inhibits the release of TSH from the anterior pituitary.</td>
<td>No. The nervous agitation is a consequence of the hypersecretion of thyroid hormone by the thyroid gland. While one will see lower levels of TSH in a patient with Graves disease, this is due to negative feedback by high circulating thyroid hormone. It is not caused by nervous agitation. Please go back and try again.</td>
</tr>
</tbody>
</table>

### Reasoning:

The next logical step is for students to establish the importance of negative feedback in maintaining thyroid hormone levels around a set point. Distractors and their associated feedback help students focus on the role of autoimmune antibodies in Graves disease and their constant and unregulated ability to stimulate thyroid hormone release. This question stimulates recall, provides guidance and feedback, and assesses performance (events 3, 5, 7, and 8).
Table 3. Continued

5. How would you link what you have said so far with the symptoms exhibited by this woman? (Choose 2.)

<table>
<thead>
<tr>
<th>Answer choices</th>
<th>Feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. The absence of negative feedback on TSH secretion has resulted in excessive stimulation of the heart and the CNS by TSH.</td>
<td>You would actually see lower levels of TSH in a patient with Graves disease, due to excessive stimulation of the thyroid gland by the antibodies that are produced in this autoimmune disease. Thyroid hormone exerts negative feedback on TSH secretion. However, there will still be excessive stimulation of the thyroid gland due to uncontrolled production of TSH-mimicking antibody. This is not a good choice; please go back and try again.</td>
</tr>
<tr>
<td>b. The antibodies produced in Graves disease have a stimulatory effect on the heart and on the CNS.</td>
<td>No. You need to talk about the effects of these antibodies on thyroid hormone secretion and then link the high levels of thyroid hormone to the symptoms displayed by this patient. Please go back and try again.</td>
</tr>
<tr>
<td>c. Excessive secretion of TSH has continually boosted metabolism, leading to increased heat production.</td>
<td>You would actually see lower levels of TSH in a patient with Graves disease, due to excessive stimulation of thyroid hormone secretion by the antibodies that are produced in this autoimmune disease. It is the thyroid hormone, not TSH, that has stimulatory effects on metabolism, and thyroid hormone levels will be very high due to the TSH-like antibody. However, the high levels of thyroid hormone will exert strong negative feedback on TSH secretion, so TSH levels will be low. Please go back and try again.</td>
</tr>
<tr>
<td>d. Uncontrolled release of thyroid hormone continually stimulates metabolism, leading to increased body temperature.</td>
<td>Yes. Because the antibody mimics TSH and because there is no negative feedback to control antibody levels, there is constant stimulation of thyroid hormone release, and the elevated levels of thyroid hormone constantly stimulate aerobic metabolism in body cells. This is one good choice. Please go back and continue choosing answers until you have two that are correct.</td>
</tr>
<tr>
<td>e. Increased heart rate and nervous agitation are a consequence of overstimulation of β-adrenergic receptors in the heart and CNS.</td>
<td>Yes. The constantly high levels of thyroid hormone will up-regulate β-adrenergic receptors, resulting in excessive stimulation of cardiac muscle cells and CNS neurons by the sympathetic nervous system. Please keep choosing answers for this question until you have two that are correct. Then continue to question 6.</td>
</tr>
</tbody>
</table>

Reasoning: This question allows students to practice applying their new basic science knowledge within a clinical context (event 6). Distractors and the associated feedback have been designed to ensure a thorough exploration of the hypothalamic–pituitary–thyroid axis, in order to enhance retention and transfer (event 9).

6. Finally, what are the treatment options for this woman and why do they have to be so drastic?

<table>
<thead>
<tr>
<th>Answer choices</th>
<th>Feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Administration of an antibody to TSH</td>
<td>No, this is not a good choice. Her TSH levels will actually be lower than normal (you should be able to explain why this is the case), and it is not TSH that is causing the hyperthyroidism. Please go back and try again.</td>
</tr>
<tr>
<td>b. Removal of the thyroid gland</td>
<td>Yes. While this seems drastic, there is no way to halt the production of these antibodies; they are not regulated by typical endocrine negative feedback. So the only option is to remove the thyroid gland. You should then talk about how this woman will need thyroid hormone replacement for the rest of her life. Well done.</td>
</tr>
<tr>
<td>c. Treatment with cortisol to suppress the immune system</td>
<td>No. This is a good thought, in that the cortisol should at least reduce the level of TSH-like antibody production by the immune system. But the immune system is important for daily ability to withstand attacks by microorganisms and to police the body for cancerous cells, so one would not want to suppress her immune system for the rest of her life. And there will still be a level of TSH-like antibody production. Please go back and try again.</td>
</tr>
<tr>
<td>d. Removal of the pituitary gland</td>
<td>No. The problem is not at the level of the pituitary gland and, indeed, her pituitary is producing less TSH than normal because of excessive negative feedback. This is not a good choice. Please go back and try again.</td>
</tr>
<tr>
<td>e. Treatment with an antibody to thyroid hormone</td>
<td>While the problem is, indeed, excessive production of thyroid hormone, this is not the approach that is used to treat Graves disease. Likely it would be excessively costly and it would be difficult to give the appropriate amount of antibody to still support normal levels of thyroid hormone. Please go back and try again.</td>
</tr>
</tbody>
</table>

Reasoning: This question, along with the feedback provided for both correct and incorrect answer choices, provides a final opportunity to stress that in Graves disease, thyroid function is no longer under negative feedback regulation (events 3 and 9).

This exercise pertains to the endocrine regulation of metabolic rate by thyroid hormone. The chart includes the SAQ, each of the six MCQs used to help students formulate an answer to that SAQ, the answer choices available for each MCQ, and the feedback for each correct and incorrect MCQ answer. The Reasoning boxes highlight the specific events of Gagné’s nine events of instruction that were applied throughout the various steps involved in the creation of this exercise.
Formative online SAQ assignments revealed no significant effect ($p > 0.05$) of the exercises on these measurements of student success (Table 4). However, while not significant ($p = 0.098$), it should be noted that there was a trend for students who had access to the formative exercises to perform better on the endocrine system SAQs in exam 1 (69.0 ± 1.40%) than the students who did not have access to the SAQ assignments (65.3 ± 1.7%). In contrast, student outcomes on the cardiovascular system SAQs (exam 2) were virtually identical ($p = 0.907$) between the two student populations.

When the focus was shifted to only the student populations provided with online SAQ exercises, it was found that the level of participation in formative assignment completion was strongly associated ($p < 0.001$) with student SAQ scores (Table 5). Student outcomes on the endocrine system (exam 1) and cardiovascular system (exam 2) SAQs were compared with their overall MCQ scores in each of these exams in an effort to isolate an effect of the formative exercises on the ability of students to formulate well-organized, complete, and accurate written answers. With regard to exam 1, the small number of students who completed only one SAQ exercise or completed neither had poorer outcomes ($p < 0.05$ and $p < 0.001$, respectively) when answering the endocrine SAQs in exam 1 compared with those students who participated fully in the online SAQ work. With regard to the MCQ outcomes, a significant difference was noted ($p < 0.001$) only when comparing the group who participated fully in the online work with those who did not participate at all. Similarly, for exam 2 and with regard to both SAQ and MCQ outcomes, significant effects of participation in the online assignments on summative examination scores were noted only when comparing those students who participated completely with those who did not attempt either assignment ($p < 0.001$ and $p < 0.05$, respectively). It should be noted when comparing student performance on the SAQ versus the MCQ portions of each of the two exams, while these values were within

### Table 4. Student outcomes on endocrine system (exam 1) or cardiovascular system (exam 2) summative SAQs during the 2 yr before (No SAQ exercises) and the 2 yr after (Access to SAQ exercises) students were provided with online SAQ assignments

<table>
<thead>
<tr>
<th></th>
<th>Exam 1</th>
<th>Exam 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n$</td>
<td>SAQ score (%)</td>
</tr>
<tr>
<td>No SAQ exercises</td>
<td>165</td>
<td>65.3 ± 1.70</td>
</tr>
<tr>
<td>Access to SAQ exercises</td>
<td>264</td>
<td>69.0 ± 1.40</td>
</tr>
</tbody>
</table>

*Mean ± SEM.

### Table 5. Influence of level of participation in online SAQ exercises on student outcomes when answering summative SAQs (SAQ score) pertaining to the endocrine system (exam 1) or the cardiovascular system (exam 2) versus exam-specific MCQs (Overall MCQ score)

<table>
<thead>
<tr>
<th></th>
<th>Number of exercises completed</th>
<th>$n$</th>
<th>SAQ score (%)</th>
<th>Overall MCQ score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exam 1</td>
<td></td>
<td>2</td>
<td>233</td>
<td>71.2 ± 1.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>21</td>
<td>58.3 ± 5.48b</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>10</td>
<td>39.8 ± 5.38c</td>
</tr>
<tr>
<td>Exam 2</td>
<td></td>
<td>2</td>
<td>216</td>
<td>74.7 ± 1.45c</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>20</td>
<td>66.8 ± 5.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>11</td>
<td>50.0 ± 6.65</td>
</tr>
</tbody>
</table>

*Mean ± SEM.

b,c Values with identical superscripts are significantly different ($p < 0.05$).

c,d Values with identical superscripts are significantly different ($p < 0.001$).
10% of each other for those students completing both assignments, student outcomes on the SAQ portion dropped to being lower \( (p < 0.05) \) than those on the MCQs by 15–23%, suggesting a stronger influence of participation in the online work on student ability to answer summative SAQs.

DISCUSSION

These interactive exercises allowed students to practice important HOCS associated with the development of well-organized answers to physiology-based questions before writing their summative examinations. While these exercises still rely on the use of MCQs, these were MCQs that prompted students to take steps in the construction of explanations pertaining to their SAQ answer by planning the answer foundation, evaluating the relevance of various pieces of physiological information, selecting appropriate information to include, and forging links between related concepts, rather than simply selecting a single answer to a fact-based question. Hence, these exercises provided opportunities for students to use HOCS such as analyze, evaluate, apply, and create as they worked their way through the series of MCQs associated with each assignment. The ability of Quandary to support the provision of a limitless number of answer options, to allow one or more correct answer choices, to either reward or penalize students based on the answer selected, and to provide immediate instructive feedback for every possible answer choice allowed the construction of interactive online exercises that guided students toward the logical organization of new knowledge, the creation of appropriate links, and the development of explanatory arguments as they practiced knowledge application within appropriate clinical and real-life contexts (Arneil and Holmes, 2009). While there are conflicting views in the literature regarding the amount of guidance that should be provided to students, ranging from purely constructivist approaches to a defined curriculum supported by didactic lectures, a strong case has been made for the provision of a reasonable level of guidance that is specifically geared toward supporting the cognitive processing needed during that particular learning process (Mayer, 2004; Kirschner et al., 2006). The provision of learner guidance is also a key instructional design principle that forms the fifth of Robert Gagné’s nine events of instruction (Gagné et al., 1992). These formative exercises also thoroughly addressed the remaining eight of Robert Gagné’s events of instruction by providing feedback-associated opportunities for students to practice both the application of their new knowledge and the organization of their reasoning when doing so. Especially worthy of note was the use of real-life clinical and physiological scenarios to capture attention, motivate learners, and provide contextual relevance. Finally, and with regard to Gagné’s seventh and ninth events, the provision of immediate instructional feedback that addressed both the correct answer and the distractors, in contrast to simply an indication of whether the answer chosen is correct or incorrect, has been shown to be important to students and to promote both knowledge retention and transfer (Moreno, 2004; Mason and Rennie, 2008; Guy et al., 2014).

The high level of student participation in these formative assignments demonstrated that students recognized the educational value of the SAQ exercises, despite the fact that each exercise contributed only 1% toward the final overall grade. The practice questions directed students to apply HOCS, referred to as generative cognitive processing, as they analyzed specific situations, applied their recently acquired physiological knowledge to an understanding of the problem to be addressed, and justified the concepts they applied or courses of action they suggested (Anderson et al., 2001; Clark and Mayer, 2008). Indeed, a strong case has been made for a role of self-generation of explanations in improving deep learning and understanding by students (Chi et al., 1994), and adult learners have frequently been described as being self-directed and learning best when conditions are task oriented (Knowles & Associates, 1984; Pratt, 1993). An additional consolidation-promoting feature of these exercises was that the final screen seen by students provided a summary of the correct answers to all the questions in that exercise, in this way providing them with a worked example of a good outline to be followed when composing a written answer. The provision of worked examples has been reported to promote learning by allowing a greater proportion of the limited capacity of short-term memory to be focused directly on the learning process itself (Kirschner et al., 2006; Clark and Mayer, 2008).

The majority of students reported that they found the SAQ exercises to be helpful, benefited from the feedback, and would like to have access to more of these formative exercises. It therefore did appear that provision of these online formative exercises improved the quality of the learning experience for these students and addressed a need to provide them with sufficient opportunities to participate in activities that would allow them to develop their critical-thinking and problem-solving skills (Crowe et al., 2008; Saunders and Gale, 2012). An important component of higher learning is the development of metacognitive skills—the ability to recognize and reflect on mistakes in order to acquire a deeper level of learning that can be applied to novel situations (Biggs, 1988; Livingston, 1997). The instantaneous feedback provided to students as they worked through the exercises, as well as opportunities to reflect and select a better answer, promoted the attainment of some important metacognitive goals (Livingston, 1997; Mynlieff et al., 2014). While about half of the respondents suggested that the exercises could have been more challenging, it is important to realize that the physiological concepts being addressed in each exercise required similar levels of problem-solving and critical-thinking skills as the SAQs these students would later encounter on their summative exams. For example, the second SAQ formative exercise asked students to identify diabetic ketoacidosis as the cause of symptoms displayed by a patient who arrived at the emergency room and to describe the physiological disruptions responsible for those symptoms, while a SAQ used repeatedly on summative exam 1 asked students to discuss the physiological cause of type 1 diabetes, certain key functions of insulin, and an important test used to monitor the management of a patient with this disease. Both questions targeted content that had been discussed thoroughly during lectures and required students to display an understanding of the key role of insulin in regulating blood sugar levels. However, it is also important to note that the MCQ-based design of the exercises, by taking students step-by-step in the formulation of their answers, would have provided some cues that could trigger information recall. The
exercises were also open-book assignments, and students were prompted to go back and try each question until it was answered correctly. These features, combined with the absence of exam-related stress or time constraints, would have promoted student well-being and achievement as they worked through the questions, leading them to perhaps think that the questions written under these conditions were easier than those encountered during summative examinations (Orr and Foster, 2013). And, most importantly, they were gaining practice in organizing and expressing their physiology-based reasoning.

Interestingly, while most students did report that they found the online assignments improved their approach to SAQs, it was not possible to demonstrate a significant effect of the introduction of these exercises on overall summative SAQ outcomes, although it should be noted that there appeared to be a nonsignificant trend toward improved outcomes for exam 1. It is frequently difficult to show a significant effect of a single modification to teaching strategy on student outcomes when average scores are compared between different years for a number of reasons, many of which apply to the current study (O’Byrne et al., 2008). Class composition changes each year, resulting in student populations with differing levels of motivation, ability, and background knowledge. While the examination questions target similar course objectives, the questions themselves do vary from year to year, and each one may not always be at precisely the same level of difficulty. Student examination schedules fluctuate, allowing the introduction of extra challenges posed by clustering of summative examinations within a short time period and the necessity for some examinations to be written in the evening, when students are tired. And, in this study, average class size increased by more than 50% just before the introduction of the SAQ exercises, and a possible confounding detrimental effect of increased class size on the learning success of some students cannot be ruled out (Lindsay and Paton-Saltzberg, 1987). However, the trend toward improved SAQ outcomes noted for exam 1 is encouraging. One would expect that any benefit gained from being guided in the development of an outline for a SAQ answer and seeing worked examples of answer outlines would be greatest with the first exposure to these exercises and that the effect would become less prominent as students became more accustomed to the process of formulating written answers (Libarkin and Ording, 2012).

In contrast, when focusing on only those students who had access to the interactive SAQ exercises, there was found to be a very strong association between the level of student participation in the online assignments and subsequent summative examination outcomes, particularly with regard to the SAQ portions of the examinations. It should be noted that completion of each exercise not only provided students with practice in problem-solving and answer-organization skills but also prompted them to reflect on course content and to consolidate their understanding of a topic when applying it in context. Hence, both learning outcomes, developing an approach to SAQ answer writing and obtaining practice in the application of new knowledge, may have contributed to the vastly improved student outcomes, especially on the SAQ portion of the examinations. But these data must also be interpreted with some caution. The majority of students completed both sets of exercises, resulting in very different sample sizes between the three populations evaluated. Furthermore, when considering the lower average exam scores for students who chose not to complete the online exercise(s), one cannot ignore a very likely inverse association between level of student participation in assignments and motivation to work hard and/or interest in course content when preparing for summative examinations. However, the positive reception of these assignments by students and the beneficial outcomes noted for students who participated fully in assignment completion suggest a value for this online work and that it should be continued and perhaps even expanded.

Finally, why put so much effort into a SAQ portion of an examination and development of student answering skills when that is so much more labor-intensive than computer-graded MCQ examinations? Preparation for an examination that will require students to write, to express their understanding of concepts and their ability to analyze and apply them in context will prompt students to use deep learning approaches (Chi et al., 1994; Scouller, 1998). Furthermore, SAQs have the ability to distinguish between students who can recognize correct answers to MCQs either through memorization of that factual information (surface learning) or a good educated guess after the elimination of some obvious distractors from those students who have acquired an in-depth understanding of that course content and can produce/synthesize their own clear explanations and justifications (Scouller, 1998; Crowe et al., 2008). For example, one study conducted by the author found that a significant proportion of students who selected the correct answer to a certain physiology-based MCQs on summative examinations were unable to correctly apply physiological principles to justify that answer choice when subsequently asked to support it with a written explanation (Carnegie, 2006).

In conclusion, online exercises that use a series of feedback-associated MCQs can be used to direct student thinking and provide practice application of HOCS when facing the challenges of teaching large classes by guiding students through the relevant concepts to be explained, linked and applied to the solution of a particular problem. The use of appropriate clinical and physiological scenarios can effectively engage student interest, and the provision of immediate feedback and opportunities to retry questions promotes reflection and an in-depth understanding of physiological concepts, their interrelationships, and their application. Similar approaches would work well for other basic sciences such as biology, chemistry, physics, and pathophysiology, because these online self-testing exercises promote scientific literacy, an important component of university-level education.

ACKNOWLEDGMENTS

This project was funded by a University of Ottawa Faculty Award for Excellence in Education to the author.

REFERENCES


Differences in Metacognitive Regulation in Introductory Biology Students: When Prompts Are Not Enough

Julie Dangremond Stanton,* Xyanthe N. Neider,† Isaura J. Gallegos,† and Nicole C. Clark‡

*Department of Cellular Biology, University of Georgia, Athens, GA 30602; †Writing Program and ‡School of Molecular Biosciences, Washington State University, Pullman, WA 99163

Submitted August 29, 2014; Revised February 13, 2015; Accepted February 24, 2015
Monitoring Editor: Debra Tomanek

Strong metacognition skills are associated with learning outcomes and student performance. Metacognition includes metacognitive knowledge—our awareness of our thinking—and metacognitive regulation—how we control our thinking to facilitate learning. In this study, we targeted metacognitive regulation by guiding students through self-evaluation assignments following the first and second exams in a large introductory biology course (n = 245). We coded these assignments for evidence of three key metacognitive-regulation skills: monitoring, evaluating, and planning. We found that nearly all students were willing to take a different approach to studying but showed varying abilities to monitor, evaluate, and plan their learning strategies. Although many students were able to outline a study plan for the second exam that could effectively address issues they identified in preparing for the first exam, only half reported that they followed their plans. Our data suggest that prompting students to use metacognitive-regulation skills is effective for some students, but others need help with metacognitive knowledge to execute the learning strategies they select. Using these results, we propose a continuum of metacognitive regulation in introductory biology students. By refining this model through further study, we aim to more effectively target metacognitive development in undergraduate biology students.

INTRODUCTION

Students who reflect on their own thinking are positioned to learn more than peers who are not metacognitive. Metacognition is a critical component of education that correlates with learning outcomes (Wang et al., 1990), student performance (Young and Fry, 2008; Vukman and Licardo, 2009), and problem-solving ability (Rickey and Stacy, 2000; Sandi-Urena et al., 2011). Owing to the significant potential to impact learning, opportunities for practicing metacognition have been included in undergraduate science courses. Posing questions to prompt students to engage in metacognitive reflection is one of the most common approaches reported in the literature (Zohar and Barzilai, 2013). For example, instructors might ask their students, “Which part of this activity was most confusing for you?” Although questions like this encourage metacognition in students, it is not clear whether all undergraduates are in a position to fully benefit from this type of prompting. Optimal targeting of metacognitive skill development requires a better understanding of metacognition in undergraduate students.

Two key elements of metacognition are metacognitive knowledge and metacognitive regulation (Brown, 1978; Jacobs and Paris, 1987). Metacognitive knowledge is our awareness of our thinking. For example, students with effective metacognitive knowledge skills can differentiate between concepts they have mastered and ones they must study further. In contrast, students lacking these skills can confuse their ability to recognize vocabulary words with mastery of the material. In undergraduate biology courses, students’ perceived knowledge may not align well with their actual knowledge (Ziegler and Montplaisir, 2014), which can
prevent them from spending more time learning the information (Pintrich, 2002). Metacognitive knowledge also encompasses an understanding of strategies for learning (Brown, 1987; Jacobs and Paris, 1987; Schraw and Moshman, 1995). This entails knowing what learning strategies exist, how to carry them out, and when and why they should be used.

While metacognitive knowledge includes the ability to identify what we do and do not know, metacognitive regulation involves the actions we take in order to learn (Sandi-Urena et al., 2011). Although the theoretical framework that delineates these components is well established in educational and cognitive psychology (Schraw, 1998; Bransford et al., 2000; Pintrich, 2002; Veenman et al., 2006; Zohar and Barzilai, 2013), biologists may not be as familiar with metacognitive regulation. Metacognitive regulation is how we control our thinking to facilitate our learning. For example, students with effective metacognitive-regulation skills can select appropriate learning strategies for a task and modify their approaches based on outcome. In contrast, students who plan to do “more of the same” after earning a poor grade on an exam lack these skills.

Metacognitive regulation is also a significant part of self-regulated learning (Zimmerman, 1986; Schraw et al., 2006). Self-regulated learners have the ability to: 1) understand what a task involves, 2) identify personal strengths and weaknesses related to the task, 3) create a plan for completing the task, 4) monitor how well the plan is working, and 5) evaluate and adjust the plan as needed (Ambrose, 2010). These abilities can form a cycle, and the last three processes (planning, monitoring, and evaluating) are key metacognitive-regulation skills (Jacobs and Paris, 1987; Schraw and Moshman, 1995; Schraw, 1998). The extent to which students use these skills is affected by their beliefs about learning and intelligence (Ambrose, 2010). For example, students who believe intelligence is fixed (Dweck and Leggett, 1988) are less likely to evaluate and adjust their plans for learning than students who believe intelligence can be developed over time and through effort.

Metacognition develops throughout the course of a person’s life (Alexander et al., 1995). Children first become aware of metacognition between the ages of three to five as “theory of the mind” develops (Flavell, 2004). At this stage, a child begins to understand that someone else’s thoughts may be different from his or her own (Lockl and Schneider, 2006). Children use nascent metacognitive-regulation skills such as planning while playing (Whitebread et al., 2009), but they do not use these abilities for academic purposes until ages eight to 10 (Veenman and Spaans, 2005). From here, metacognitive regulation grows linearly through middle and high school (Veenman et al., 2004) and is thought to advance well into adulthood (Kuhn, 2000; Vukman, 2005). Therefore, metacognition is likely to be an area of ongoing development for young adults. While we know that awareness and control of thinking progress over time, we do not know the important stages that occur in the metacognitive development of undergraduate students (Dinsmore et al., 2008; Zohar and Barzilai, 2013). To enhance student learning through improved metacognition, we need to characterize the key transitions that occur as undergraduates acquire metacognitive-regulation skills. As a first step toward identifying these transitions, we asked what metacognitive-regulation skills are evident in undergraduates taking an introductory biology course. We used the task of preparing for an exam as a vehicle for examining metacognition. We reasoned that students would see this as an important endeavor that would merit their reflection. Our primary interest was in the metacognitive-regulation skills students used while preparing for an exam rather than the particular study strategies they selected. To this end, students were given a self-evaluation assignment after the first exam in the course and a follow-through assignment after the second exam. We used qualitative methods to analyze assignments for student statements of monitoring, evaluating, and planning. On the basis of these data, we propose a continuum of metacognitive-regulation development in introductory biology students. We use this working model to generate hypotheses for further study and to make suggestions for instructors interested in facilitating student metacognition.

METHODOLOGY

Participants and Context

Participants were students in an introductory biology lecture and lab course at a public land-grant university, with an RU/VH Carnegie Foundation classification (research university with very high research activity). BIOL107, Introductory Biology, focuses on cell biology and genetics, and is one part of a yearlong introductory series that can be taken in any order. One semester of college chemistry is the only prerequisite. The course serves ~350–450 students each semester through one lecture section taught by a single professor and 18–22 lab sections taught by graduate teaching assistants (TAs). The majority of the students are freshmen and sophomores, but juniors and seniors also take the course. Although BIOL107 is intended for science and pre-professional majors, some nonmajors take it as a general education requirement.

All BIOL107 students were given the metacognition assignments described below, because one of the course goals is for students “to develop an approach to the study of biology that will facilitate success in future courses.” Only students who were 18 years or older, gave written informed consent, and completed both assignments were included in this study. In the Fall 2013 semester, 245 of the 346 students who completed BIOL107 met these criteria (n = 245 for this study). One student was not included because he was a minor, 18 students did not sign the consent form, and the remaining nonparticipating students did not turn in one or both of the assignments (described in the following section). Each assignment was worth five points or 0.7% of the total course grade (for a total of 10 points or 1.4%), and this may not have been enough of an incentive for students to complete both assignments. The Washington State University Institutional Review Board declared this study exempt (IRB 12702).

Metacognition Assignments

After the first exam in the course, students were given a self-evaluation for exam 1 assignment (E1-SE, see Supplemental Material, Appendix 1). This two-page assignment included three short-answer and six open-ended questions/prompts written to encourage students to monitor
evaluate the strategies they used to prepare for exam 1 and to create a study plan for exam 2. As part of the assignment, students read a list of study strategies used by students who earned a grade of “A” on exams in previous semesters of the course (see Supplemental Material, Appendix 3). This list included all of the methods provided by the “A” students; more “passive” approaches were not filtered out. On E1-SE, students were asked to consider including one or more of these strategies in their study plans for exam 2. This assignment was piloted in the Spring 2013 semester and modified slightly following initial data analysis.

A second assignment was created based on preliminary data collected from Spring 2013, which indicated that nearly all students were willing to modify their study plans in some way. Following the second exam, students were given an exam 2 follow-through assignment (E2-FT, see Supplemental Material, Appendix 2) designed to determine whether they spent more time studying for exam 2, and the extent to which they followed their study plan. This one-page assignment was composed of two short-answer and three open-ended questions/prompts, and it also asked students to create a study plan for exam three. Students had copies of their completed E1-SE assignment when answering these questions. They were given 1 wk to complete E1-SE and E2-FT, and graduate TAs collected the assignments. Students earned full credit for completing each assignment (five points each, 10 points total); points were not awarded based on the quality of their answers. Together, E1-SE and E2-FT constituted only 1.4% of the total course grade.

Qualitative Data Analysis

E1-SE. Students’ metacognition assignments were studied using content analysis. In essence, we used students’ words to make inferences about their levels of metacognitive regulation. First, we read all of the E1-SE assignments with metacognitive regulation in mind. Initially, we used magnitude codes (Weston et al., 2001; Saldaña, 2013) to score students’ statements about metacognitive regulation as high, medium, or low; however, it was difficult for us to distinguish among the three levels. Instead, we developed a coding system in which we rated the students’ statements as either “sufficient/provides evidence” or “insufficient/provides no evidence” (Tables 1 and 2). These ratings were assigned to each student for: 1) monitoring exam 1 strategies and 2) evaluating exam 1 strategies when planning for exam 2. We coded evaluating and planning together, because we did not specifically ask them to rate their exam 1 study plan. Instead, we asked them to select learning strategies for exam 2 (planning) using insights from their exam 1 experiences (evaluating). We also coded direct student quotes that exemplified the range of metacognitive regulation students reported using, and we coded statements that described any actions the students reported taking (Saldaña, 2013). Two authors (J.D.S. and N.C.C.) coded ~10% of the assignments before discussing them for approximately 2 h and developing an initial codebook. Another 10% of the assignments were coded and discussed, and the codebook was revised. Once the codebook was fully developed, we coded the remaining assignments in ~20% increments, with discussions taking place after each increment. The process was iterative, and assignments were recoded whenever the codebook was revised.

The final codes for each assignment were compared, and in cases of disagreement, we discussed them until we were able to come to consensus.

In qualitative research, the researchers are the instruments, and they act as filters through which the data are analyzed (Bogdan and Biklen, 2003; Denzin and Lincoln, 2003; Yin, 2010). In this study, the researchers who coded came to the data with different perspectives: as an instructor (J.D.S.), as a graduate TA (I.J.G.), and as an undergraduate student (N.C.C.). We found that each of us could bring insights to the data. Guided and checked by a researcher with extensive training and experience in qualitative methods (X.N.N.), we completed multiple cycles of coding the data separately and then discussing the data together as described earlier. This approach allowed us to discover nuanced details that would otherwise be overlooked if our primary goal was to calculate interrater reliability (Bogdan and Biklen, 2003; Denzin and Lincoln, 2003, 2005).

E2-FT. Two authors (J.D.S. and I.J.G.) used the same content analysis process described above to code E2-FT. We noted when students reported that they spent more time studying for exam 2 than exam 1. We also coded for evidence that they followed the key features of the study plans for exam 2 that they outlined in E1-SE. Student statements were coded as “yes/followed plan” or “no/did not follow plan.” We gave students the benefit of the doubt when we were not certain, because we did not have a way to determine whether the students spent more time or followed their plans. For example, if a student said he or she followed the plan and provided no information to the contrary, we coded the E2-FT with “yes/followed plan.”

RESULTS

To study the extent to which metacognitive-regulation skills were evident in introductory biology students, we examined two metacognition assignments related to exam preparation. As described in the Methods, the E1-SE was given after exam 1, and the E2-FT was given after exam 2. Owing to our interest in understanding metacognitive regulation, we focused on evidence of students’ monitoring, evaluating, and planning of learning strategies rather than the learning strategies themselves.

Monitoring Exam 1 Learning Strategies

We examined students’ ability to monitor the effectiveness of their learning strategies. This skill should not be confused with monitoring of conceptual understanding, which is another important metacognitive skill (see Evaluating Exam 1 Study Plan While Planning for Exam 2). After describing their approaches to studying for exam 1, students were asked to respond to two prompts: 1) “Now that I have seen the grade I earned on exam one, these are the study strategies that I feel worked well for me, and I plan on using them again for exam two” and 2) “Now that I have seen the grade I earned on exam one, these are the study strategies that I feel did not work well for me, and I don’t plan on using them again for exam two.” We found evidence of strategy monitoring in 120 of 245 assignments (49.0%; Table 1). For example, two students referred to the use of note
cards. One student explained why note cards were helpful for her, while another student explained why flashcards were not effective for her.

In response to what worked: “Doing note cards made me read through the material and put it into my own words so that I could understand it.”

In response to what did not work: “I don’t feel that my flashcard method worked because I focused on terms instead of concepts.”

As another example of monitoring learning strategies, several students explained why “looking over,” “going over,” or “skimming” course materials such as the textbook, class notes, and online homework assignments were not effective strategies:

In response to what did not work: “I feel just looking at lecture slides did not help me. I felt that time would be better spent to quiz myself and see what I don’t know.”

In response to what did not work: “When I would look at notes and homework and not write anything down to study from. I need to be interactive with the material to fully grasp the concepts.”

Table 1. Qualitative analysis of monitoring learning-strategy effectiveness for exam 1 (E1-SE)

<table>
<thead>
<tr>
<th>Monitoring code</th>
<th>Percentage of students</th>
<th>Example student response and content analysis notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sufficient evidence</td>
<td>49.0 (120/245)</td>
<td>Quote: “I don’t feel that my flashcard method worked because I focused on terms instead of concepts.” Notes: student identifies a learning strategy that was not effective and provides a specific explanation for why it was not effective.</td>
</tr>
<tr>
<td>Insufficient evidence</td>
<td>51.0 (125/245)</td>
<td>Quote: “It’s not necessarily study strategies didn’t work for me, it’s that I didn’t try enough. I need to try more!” Notes: student does not identify any learning strategies that were effective or ineffective and provides an ambiguous reason for exam results.</td>
</tr>
</tbody>
</table>

“To examine monitoring of learning-strategy effectiveness, we asked students to respond to two prompts focused on the approaches that worked and did not work for exam 1 (see Results). Using content analysis, we coded students’ responses as providing sufficient or insufficient evidence of monitoring. The percentage and number of students in each category are shown (n = 245).

Table 2. Qualitative analysis of evaluating and planning for exam 2 (E1-SE)

<table>
<thead>
<tr>
<th>Evaluating and planning code</th>
<th>Percentage of students</th>
<th>Example student response and content analysis notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sufficient evidence (willing to change)</td>
<td>44.9 (110/245)</td>
<td>New strategy selected: reviewing notes after each class Quote: “This will allow me to focus on areas where I am struggling and ask questions sooner.” Notes: student is willing to change his study plan for exam 2. He selects a new strategy and provides a reason based on his exam 1 experience.</td>
</tr>
<tr>
<td>Insufficient evidence (willing to change)</td>
<td>53.5 (131/245)</td>
<td>New strategies selected: pursue free tutoring on campus and join a study group Quote: “Anything can help if you have not yet received a hundred percent.” Notes: student is willing to change his study plan for exam 2, however, he selects two new strategies without providing any reasons based on his exam 1 experience.</td>
</tr>
<tr>
<td>Insufficient evidence (unwilling to change)</td>
<td>1.6 (4/245)</td>
<td>New strategy selected: not applicable Quote: “I got 78% with NO studying. I think I will be ok. I’ve taken tests before.” Notes: student is not willing to change her study plan for exam 2 and is not reflecting on her exam 1 experience.</td>
</tr>
</tbody>
</table>

“We posed two prompts and one question to assess whether students reflected on their exam 1 experiences and adjusted their study plans for exam 2 accordingly (see Results). We used content analysis to code student responses as providing sufficient or insufficient evidence of evaluating and planning (n = 245). A small percentage of students (1.6%) reported that they would not do anything differently for exam 2. These students are shown in the bottom category “Insufficient evidence/unwilling to change.”

but did not provide any explanation for why a particular strategy did or did not work for them (Table 1). Within the group that did not provide evidence of monitoring, an important theme emerged. Seventy-five students expressed that they did not need to use different learning strategies; they only needed to spend more time studying.

In response to what worked: “I feel my strategies were sound I just need to begin studying earlier.”

In response to what did not work: “I must start studying earlier! I think my strategies were good, but my time spent was poor.”

While 120 of 245 students provided evidence that they monitored the effectiveness of individual learning strategies, the other 125 students (51.0%) listed their approaches
In a few cases, students reported that the strategies they used for exam 1 did not work, but they still included them in their plans for exam 2. These students seemed to know that they should try something different but did not do so for reasons that could not be discerned.

In response to what worked: “After seeing my grade, these study strategies obviously were not helpful because I didn’t do well at all.”

This student did not provide evidence of monitoring of her exam 1 learning strategies. Interestingly, in the same assignment, the same student said,

In response to what didn’t work: “I plan on using all the same things but adding other strategies on top of that.”

Related to inability to monitor learning-strategy effectiveness, another sentiment was some students’ inability to monitor their exam performance. Sixteen students specifically described this problem.

“I felt like I did great on (exam one) but my grade doesn’t reflect it.”

“I feel like nothing I did worked because I felt very confident before and after the exam; however, I ended up getting a 65%.”

The difference between how students predict they performed on an exam and how they actually performed is directly related to metacognition (Pieschl, 2009; Ziegler and Montplaisir, 2014). Students lacking metacognitive knowledge cannot accurately judge what they know and, more importantly, what they do not know. This inability can be a barrier to learning. For example, if students believe they have mastered concepts because they recognize terms, this may cause them to underprepare for an exam and then overestimate their performance.

In summary, we found evidence of learning-strategy monitoring from 49.0% of the students. These students could explain why an approach they took was successful for them or not. The other 51.0% of the students did not provide evidence that they were monitoring the effectiveness of individual strategies they used. In many cases, students pointed to time rather than approach as being the most important factor.

**Evaluating Exam 1 Study Plan While Planning for Exam 2**

We wanted to know whether students at this level could reflect on their study plans for exam 1 and adjust them based on their performance. We asked students to respond to two statements: 1) “A compiled list of study strategies used by students who earned high grades on Biology 107 exam 1 in past semesters is posted (online). After reading this document, I might try the following new study strategy for exam two” and 2) “The reason I think this may be helpful is.” Next, we asked students: “Besides what you already wrote, what else do you plan to do differently for exam two now that you have the experience of taking exam one?” Surprisingly, 241 out of 245 students (98.4%) were willing to change their initial study plan. Only four out of 245 (1.6%) reported that they would not change their plans in any way. One of these four students stated this was because she earned 100% on exam 1. While almost all students selected at least one new learning strategy, only 110 students out of 245 (44.9%) indicated that their selection was based on their exam 1 experiences (Table 2).

New strategy selected: “Write detailed answers to the study questions every week.”

In response to why this would be helpful: “It will make studying for the exam easier because I will already understand the key concepts that I might only need to briefly review later rather than ‘study.’”

New strategy selected: “Take my own detailed notes.”

In response to why this would be helpful: “I will retain the information in a way that I understand rather than copying down the notes word for word.”

Forty students recognized that a new strategy they selected could help them monitor conceptual understanding (a skill different from but related to monitoring effectiveness of learning strategies), so that they could identify concepts they had not yet mastered.

New strategy selected: self-testing

In response to why this would be helpful: “I think (this technique) will help me because just simply running over the material is not enough. I need to practice without looking at notes to see what I know and what I need to spend more time on.”

New strategy selected: review notes after each class

In response to why this would be helpful: “This will allow me to focus on areas where I am struggling and ask questions sooner.”

The remaining students, 135 out of 245 (55.1%), did not provide evidence that they were reflecting on and adjusting their plans based on exam 1 performance (Table 2). Often, these students did not seem to know why a new strategy might be effective.

New strategies listed: review notes weekly, review study questions, and take the practice test

In response to why these would be helpful: “These are all things I didn’t do and got a poor grade, if I do what people did that got an A it can only help.”

New strategies listed: pursue free tutoring on campus and join a study group
In response to why a new strategy would be helpful:
“Anything can help if you have not yet received a hundred percent.”

In summary, very few students were unwilling to do something different for exam 2. We found that 44.9% of students could evaluate their study plans for exam 1 and adjust their plans for exam 2 accordingly. A theme that emerged from these students was monitoring of conceptual understanding. Among the other 53.5% of students who changed their plans for exam 2 were several students who did not describe why a new strategy might be helpful to them.

**Student Realizations about Studying after Exam 1**
While coding E1-SE, we found two common realizations made by students in the process of reflecting on exam 1. Both are related to the self-regulated learning skill of understanding the task. The first theme was the realization that engagement with the material is required. Thirty-three students (13.5%) recognized that success in the course would require active approaches to learning.

In response to what did not work: “Just attending class and doing the homework is not enough.”

In response to what did not work: “Only studying the content and not applying them [sic] to understand completely. Studying concepts individually and not connecting them to make sense.”

The second theme was the realization that more exposure to the material is required. More than half of the students (55.5%) saw the need to spend more time with the material, yet many still focused on retention of information.

“I need to not just cram for the exam, if I can at least grasp all the info before an exam by studying more my (study) sessions would be more a review than trying to learn everything over a weekend.”

“Your brain cannot remember information over a two day period. Spreading the time out allows for less stress and more memorization and understanding.”

Both of these realizations could further metacognitive regulation, because not only did the students recognize that their study plans were not effective, but they were also beginning to understand why they were not effective. This understanding may have helped them select strategies aimed at increasing engagement and exposure. Additionally, these realizations relate to the self-regulated learning skill of being able to assess the assignment. In this case, students were beginning to understand the nature of the exams in the course and what these assessments might require of them.

**Time Spent Studying for Exam 2**
Because the need to devote more time to studying was a common statement on the exam 1 self-evaluation, we asked students if they spent more time preparing for exam 2 than they did for exam 1. On the E2-FT, 183 of 245 students (74.7%) said they spent more time studying for exam 2. When asked how they were able to spend more time, students reported that they: started earlier, scheduled time to study, and spread their studying out over time.

“Instead of only doing work for one class for an extended period of time I would switch between classes so that I studied each over a longer period of time.”

Some students mentioned that they were motivated to improve their grades, which allowed them to make studying a priority.

“I was able to put more time in because I made more time for studying and cut off some free time I was doing nothing with.”

Sixty-two students (25.3%) said they did not spend more time. While a few mentioned this was because they did not need to, most of these students gave other reasons. Common factors included: work, illness, other classes, and factors outside of school. Some indicated that they did not make studying a priority.

In response to why more time was not spent: “Time escaped me.”

“[I] didn’t have much time and didn’t get around to it when I did have time.”

Following the experience of exam 1, most students were able to spend more time studying for exam 2. We were curious to know whether following a new study plan would be as achievable.

**Study Plan for Exam 2**
We were surprised that 241 students (98.4%) reported on E1-SE that they would change their exam 1 study plans to prepare for exam 2. On E2-FT, we asked students whether they followed their study plans for exam 2, and if so, how they were able to do so. We coded for evidence that they used the learning strategies they outlined in the E1-SE assignment. We found that 125 out of 245 students (51.0%) followed key parts of their plans. Although students were generally unable to explain the mechanism that enabled them to change, some described a commitment to studying as an important factor.

“I made myself follow my study plan. I went to the library and told myself that I could not leave until I had studied for two hours, not counting breaks in concentration. I did this every day leading up to the test starting the Friday before.”

“I actually sat down and studied, while on exam one, I never actually studied because I thought I knew the material.”

Other students described a mind-set or motivation that may have allowed them to try new learning strategies:
“I felt that by changing some of the ways I study it would help me more than just studying longer.”

“I was not happy about the first exam’s result so I was motivated to increase my grade. Thus, I studied more materials.”

The other 49.0% (120 out of 245 students) did not provide evidence that they followed their study plans for exam 2. When asked why they were not able to, some students explained that they thought they did not need to.

“Although I planned to read over the questions and do the study questions every week, instead I assumed it because I could explain it but I struggled applying it.”

Some students reverted to using learning strategies for exam 2 that they previously said did not work for them on exam 1. Other students selected more active approaches to studying on E1-SE and then replaced them with more passive approaches. For example, on E1-SE, one student planned to test his understanding of the material by taking the practice test without notes and answering the study questions without notes for exam 2, but reported on E2-FT that he decided to watch online biology videos instead. Other students gave reasons why their plans were not easy to follow.

“I didn’t read as much of the textbook as I would have liked. It’s super painful to read.”

“I studied a little more but if I study too far in advance, I can’t remember the material I studied.”

Data from E2-FT suggest that while spending more time studying was doable for most students, trying a new learning strategy was not as easy to carry out. Most students were willing to change, but because it is difficult to change, students may need more help in order to do this.

**DISCUSSION**

We studied metacognition in undergraduates by using exam preparation as a mechanism for investigating their monitoring, evaluating, and planning skills. We found evidence that approximately half of the students (49.0%) monitored the effectiveness of the learning strategies they used for exam 1 (Table 1). These students could identify strategies that were helpful and unhelpful, and they provided explanations for their answers. While monitoring, some students gained a better understanding of the task, which is another important self-regulated learning skill (Ambrose, 2010; Meijer et al., 2012). Students reported that more engagement with the material was required (13.5%), and greater exposure to the concepts was necessary (55.5%). These realizations are valuable, because they prime students to use metacognitive-regulation skills such as planning more effectively. The rest of the students (51.0%) were not able to monitor the effectiveness of their learning strategies and usually wrote ambiguous statements to explain why all or none of their approaches worked. This is not surprising, given that monitoring is a skill that develops later in life than evaluating and planning (Schraw, 1998), and it can be weak even in adults (Pressley and Ghatatala, 1990; Alexander et al., 1995).

Although the monitoring data were interesting, what was most intriguing was the fact that nearly all of the students (98.4%) were willing to select new learning strategies as part of their exam 2 study plans on the E1-SE. As researchers who are invested in teaching, we found this openness to change encouraging. It suggests that the students were beginning to reflect on their approach to studying, which is a first step toward using metacognition to regulate learning. While analyzing student responses for evidence of evaluating and planning, we found that 53.5% of the students did not seem to know which new strategies would be appropriate, despite a willingness to change (Table 2). They selected strategies from the list provided, but they could not give a reason why alternative approaches might be helpful to them (Table 2, Evaluating and planning insufficient evidence/willing to change). Conversely, we were impressed by the percentage of students (44.9%) who could reflect on their first study plans and select strategies based on their exam 1 experiences (Table 2, Evaluating and planning sufficient evidence).

Because metacognitive regulation involves the actions we take to learn (Sandi-Urena et al., 2011), we wanted to know whether students who reflected on exam 1 and adjusted their plans for exam 2 carried out the new plans they made. As part of the E2-FT, we asked whether or not students followed their study plans. We found that only half of the students reported following the key parts of their plans that related to their exam 1 experience. When the other half was asked why they did not follow their plans, these students explained that they did not need to change and/or they did not know how to change. This fits with a previous study that showed introductory biology students do not use active-learning strategies if their courses do not require them to (Stanger-Hall, 2012), and they will not use deep approaches to learning if they do not know how to (Tomanek and Montplaisir, 2004). We conclude that prompting students to use metacognitive-regulation skills is enough for some students to take action, but others need additional instruction in order to respond optimally (Zohar and Barzilai, 2013).

Using our data, we have outlined a working model to represent potential categories of metacognitive-regulation development represented in this population of undergraduates. We suggest four possible metacognitive-regulation categories: “not engaging,” “struggling,” “emerging,” and “developing” (Table 3 and Figure 1), which we propose exist in a continuum. We then use the data to generate hypotheses, make predictions, and provide suggestions for instructors.

“Not Engaging” in Metacognitive Regulation

Very few students were unwilling to change any parts of their study plans. We placed these students in a category in the continuum that we describe as “not engaging” in metacognitive regulation (Table 3). They saw no reason to alter their approach to studying, and therefore did not select new learning strategies. Their plans for exam 2 remained the same as for exam 1. Although we did not set out to study agency and self-efficacy, these ideas appeared in the students’ assignments. Briefly, agency refers to the belief that learning is your responsibility (Baxter Magolda, 2000), while
in a dualism position believe that the instructor knows everything and that it is the instructor’s job to teach students the correct answers (Perry, 1968; Markwell and Courtney, 2006). If this hypothesis holds true, then we predict that lack of agency will prevent “not engaging” students from recognizing the need to use metacognitive-regulation skills. For example, if a student believes that instructors determine student performance, then he or she is unlikely to focus on monitoring, evaluating, and planning his or her strategies for studying.

Hypothesis: Students “Not Engaging” in Metacognitive Regulation Are Unable to Monitor Their Conceptual Understanding of Material

We also hypothesize that “not engaging” students struggle with the ability to accurately assess what they do and do not know. If this holds true, then we predict that inability to monitor conceptual understanding will be another barrier to metacognitive regulation. For example, if a student believes that instructors determine student performance, then he or she is unlikely to focus on monitoring, evaluating, and planning his or her strategies for studying.

Hypothesis: Students “Not Engaging” in Metacognitive Regulation Lack Agency

We hypothesize that students in the “not engaging” category lack agency and do not believe they can determine their own success in a course. They may also have a dualistic view, seeing the world in “black and white,” where there is only right or wrong (Perry, 1968; Markwell and Courtney, 2006). Students

Proposed continuum in introductory biology students

<table>
<thead>
<tr>
<th>Metacognitive category:</th>
<th>Not engaging (rare)</th>
<th>Struggling</th>
<th>Emerging</th>
<th>Developing (rare)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluating and planning</td>
<td>Insufficient evidence of evaluating</td>
<td>Insufficient evidence of evaluating</td>
<td>Sufficient evidence of evaluating</td>
<td>Sufficient evidence of evaluating</td>
</tr>
<tr>
<td>Followed study plan</td>
<td>Did not change study plan for exam 2</td>
<td>Some followed study plan for exam 2</td>
<td>Carrying out appropriate strategies</td>
<td>Not identified</td>
</tr>
<tr>
<td>Metacognitive challenge</td>
<td>Recognizing the need to use different study strategies</td>
<td>“I am going to do the same things but work harder ... because I apparently cannot figure out what you are asking.”</td>
<td>“It is a different way for me to study and understand biology. It is out of the norm of the way I study so maybe it will help.”</td>
<td>“[The weekly questions] effectively enforce thinking of how concepts are tied together and how they apply to the real world.”</td>
</tr>
<tr>
<td>Example quotes in response to why a new strategy would be helpful</td>
<td>“I clearly could not apply the knowledge to the type of questions you asked. Rather than straight forward questions that I can confidently answer correctly, these are curveball questions that do not test my knowledge but how well I can interpret the meaning of the question.”</td>
<td>Students may need help developing metacognitive knowledge and metacognitive-regulation skills.</td>
<td>Students may need help developing procedural knowledge and metacognitive-regulation skills.</td>
<td>Prompting may be enough to encourage optimal metacognitive regulation.</td>
</tr>
<tr>
<td>Notes for instructors</td>
<td>Students may need help with beliefs about learning and monitoring conceptual understanding.</td>
<td>Students may need help developing metacognitive knowledge and metacognitive-regulation skills.</td>
<td>Students may need help developing procedural knowledge and metacognitive-regulation skills.</td>
<td>Notes for instructors</td>
</tr>
</tbody>
</table>

*Using data from the E1-SE and the E2-FT assignments, we propose a continuum of metacognitive regulation in introductory students and make suggestions for instructors on how to help students in each category.*

**Self-efficacy** is the belief that you are capable of learning (Bandura, 1997; Estrada-Hollenbeck et al., 2011; Trujillo and Tanner, 2014). Both agency and self-efficacy have the potential to affect students’ metacognitive regulation. We found evidence that students in the “not engaging” category felt they were capable of learning but did not necessarily see it as their responsibility:

“I clearly could not apply the knowledge to the type of questions you asked. Rather than straight forward questions that I can confidently answer correctly, these are curveball questions that do not test my knowledge but how well I can interpret the meaning of the question.”

**Hypothesis: Students “Not Engaging” in Metacognitive Regulation Lack Agency**

We hypothesize that students in the “not engaging” category lack agency and do not believe they can determine their own success in a course. They may also have a dualistic view, seeing the world in “black and white,” where there is only right or wrong (Perry, 1968; Markwell and Courtney, 2006). Students
In addressing the “not engaging” category, we recommend that instructors include formative assessments that help students identify concepts they do not truly understand. For example, regular online quizzes with immediate feedback on incorrect answers could help students in the “not engaging” category confront their misconceptions.

“Struggling” with Metacognitive Regulation

Our data indicate that most of the students in this study fit into categories we describe as “struggling” with or “emerging” metacognitive regulation. Students in the “struggling” category (Table 3) were willing to change their study plans, but they often used noncommittal language such as “might” and “try” regarding new learning strategies. They did not choose strategies that addressed issues they reported having. For example, a student wrote on his E1-SE that he could narrow down the answers to multiple-choice questions to the two most likely options, but he did not have the depth of knowledge to select the correct one. Yet this student’s exam 2 study plan centered on reading chapter summaries, which was not likely to give him the level of understanding he needed. Students in the “struggling” category also selected passive approaches that involved “going over” and “looking over” course material. This is consistent with recent research showing that undergraduate chemistry students primarily focus on reviewing material when studying (Lopez et al., 2013). “Struggling” students seemed to have agency, but they lacked self-efficacy and were not confident in their ability to select appropriate learning strategies. Specifically, students in the “struggling” category indicated that they were willing to modify their study plan, but some reported that they did not know what to change.

“Honestly I don’t know what to do. Even starting to study a week in advance, I got a lower score then when I (didn’t study) at all. I do not know what works for me in this class.”

Hypothesis: Students “Struggling” with Metacognition Lack Metacognitive Knowledge

We hypothesize that students in the “struggling” with metacognitive-regulation category lack metacognitive knowledge. Metacognitive knowledge can be divided into declarative knowledge, procedural knowledge, and conditional knowledge (Brown, 1987; Jacobs and Paris, 1987; Schraw and Moshman, 1995). Declarative knowledge includes knowing about yourself as a learner, procedural knowledge involves knowing what learning strategies exist and how to use them, and conditional knowledge entails knowing when and why to use a learning strategy (Schraw and Moshman, 1995). In this study, we provided students with a list of study approaches used by past students who earned a grade of “A” on exams in the course. This could have helped students become aware of existing strategies, but it would not help them with procedural and conditional knowledge if they did not know how, when, and why to utilize an approach they picked from the list. If “struggling” students lack metacognitive knowledge, then we predict that they can move to the “emerging” category (see “Emerging” Metacognitive Regulation) if they are given direct instruction in learning strategies.

To help students develop metacognitive knowledge, we recommend that instructors create class activities that introduce students to learning strategies and help students understand how to execute those strategies (procedural knowledge). Strategies should be explored in the context of course material (Dignath and Büttner, 2008), so students can begin to recognize when and why to use them (conditional knowledge; Meijer et al., 2012). For example, an instructor could explain what a concept map is and what it is used for (Novak, 1990) and then model this approach using a recent course topic while explaining his or her thought processes (Schraw, 1998). This think-aloud technique gives students insights into the metacognition the instructor uses while carrying out the strategy (Kolencik and Hillwig, 2011). Following this demonstration, the instructor could ask students to generate a list of topics that are best served by a concept map and then facilitate a class discussion of student ideas. This type of approach could help students in the “struggling” category gain metacognitive knowledge, which would help them select learning strategies that better align with their studying needs.

“Emerging” Metacognitive Regulation

Other students in our study seemed to be in a later part of the continuum, and they could be characterized as possessing “emerging” metacognitive regulation (Table 3). These students recognized a need to change their study plans for exam 2, but in contrast to “struggling” students, those with “emerging” metacognitive regulation could select appropriate learning strategies. For example, a student wrote that she struggled to make connections between concepts on exam 1 because she focused on terms while studying. For exam 2, she planned to draw diagrams to help identify relationships between concepts. “Emerging” students also recognized the importance of trying to understand the material rather than just retain it. Students in the “emerging” category did not always follow their study plans but seemed to have both agency and self-efficacy.

Example of agency: “[The exam] was a little more difficult than I expected, I’m sure that this is because I wasn’t as prepared as I should/could have been.”

Example of self-efficacy in response to why she would continue using a learning strategy: “I’m able to understand the concepts more and by doing this I can sometimes find answers to my own questions about a topic or subject.”

Hypothesis: Students with “Emerging” Metacognitive Regulation Have Conditional Knowledge, but Lack Procedural Knowledge

Students in this category were willing to change their study plans and could select appropriate learning strategies but did not always carry them out. We hypothesize that “emerging” students lack metacognitive knowledge and, specifically, procedural knowledge of how to use approaches to learning. Whereas “struggling” students also lacked both procedural and conditional knowledge, our data suggest that “emerging” students know what learning strategies exist and when and why to use them. We predict that “emerging” students...
can move to the “developing” category (see “Developing” Metacognitive Regulation) of metacognitive regulation if they are provided with training in procedural knowledge.

To help students with procedural knowledge, we suggest that instructors follow the three steps for metacognitive skill development outlined by Veenman et al. (2006). First, instructors should model learning strategies using relevant course topics, as recommended for “struggling” students. It is also valuable to have an experienced student model the learning strategy, because a peer’s ability to use the strategy will be closer to that of the “emerging” students (Schraw et al., 2006). This will not only help students better understand the steps involved in a learning strategy, but it will also make it more difficult for them to claim that a particular approach is not doable (Bandura, 1997). Second, instructors should also be very explicit about the benefits of each learning strategy, which may help students overcome the perceived difficulty involved in trying something new. Third, it is recommended that instructors train students over time and provide ample opportunity for practice (Tomanek and Montplaisir, 2004), so students can fully understand how to use the learning strategies. Students will also benefit from instructor feedback as they try to acquire new metacognitive knowledge skills (Schraw et al., 2006). Further development of procedural knowledge could help “emerging” students move to the next category of metacognitive regulation.

“Developing” Metacognitive Regulation

We found a small percentage of students in a “developing” metacognitive-regulation category (Table 3). These students recognized the benefit of adjusting their study plans and could select learning strategies for exam 2 that appropriately addressed issues they had in preparing for exam 1. They reported on E2-FT that they had followed their study plans for exam 2. Interestingly, “developing” students who earned high grades on exam 1 still used their metacognitive-regulation skills to enhance their learning. For example, a student who scored 100% on exam 1 identified an ineffective strategy that he would no longer use (flashcards) and selected new strategies (creating visual representations of concepts and integrating notes from different sources) with the goal of refining his study plan. Another student who earned a high grade reported that his exam 1 study plan was effective, but he still wanted to adjust his approach based on what he learned about exams in the course.

“I will spend more time thinking about how the concepts can be applied to real-world situations; there were a lot more critical-thinking questions on the exam than simple facts.”

“Developing” students focused on studying to learn rather than to earn high grades. In one theoretical framework, this goal is associated with a mastery approach to learning and is contrasted with a performance approach focused on showing competence (Ames and Archer, 1988; Heyman and Dweck, 1992). In another framework, this goal is categorized as a deep approach to learning, with a focus on meaning rather than memorization (Biggs, 1987). While other students reported that they selected strategies because those strategies would help them memorize material so they could do well on the exam, “developing” students focused on understanding the concepts. For example, a student explained why it would be helpful to answer study questions without her notes.

“It requires that I fully understand on my own and can explain it. I will know I understand the information if I can say what it is without needing to ‘jog my memory.’”

Students in this category also made statements that demonstrated their agency and self-efficacy. Self-efficacy in undergraduate students strongly correlates with metacognition (Coutinho and Neuman, 2008). Students who believe they are capable of learning are likely to use metacognition to improve their understanding.

“Some days, I get behind and think I’ll learn it later on. I know that the concepts in this course build upon each other so I should take the time to understand main points for each topic.”

Prompts such as the questions posed in our postexam assignments can be effective in encouraging “developing” students to engage in metacognitive regulation. This fits with data indicating that a mastery approach to learning is correlated with use of metacognition (Coutinho and Neuman, 2008). We hypothesize that more “developing” students might be found in upper-division biology courses. We are currently studying students in 300- and 400-level biology courses to further document metacognition in “developing” students. We plan to use this data to train other students to emulate “developing” students’ use of metacognitive-regulation skills.

Limitations of Study/Alternative Explanations

We gained several insights into the metacognitive regulation used by introductory biology students through metacognition assignments. Nevertheless, it is important to acknowledge that these are “offline assessments” occurring before or after a metacognitive event (van Hout-Wolters, 2000; Veenman et al., 2006). This type of assessment relies on a student’s ability to accurately remember and report what he or she did. For example, as noted in the Methods, we did not know whether students actually followed their plans or not. We looked at whether or not they checked “yes” when asked if they followed their plans, and we also considered whether the students provided any evidence to the contrary. In addition to the limitation of self-report, written data do not allow researchers to ask participants follow-up questions to clarify their intended meaning. To address these concerns, in our current studies, we are interviewing students about their studying, and we will use think-aloud protocols to do “online assessment” of metacognition as it is happening (Meijer et al., 2012).

We were surprised that so few students at the introductory level fit into the “not engaging” category of metacognitive regulation based on their postexam assignments. It is possible that students felt pressure to write what they thought the instructor wanted or that they may have had social desirability bias, the desire to show themselves in a favorable light (Gonyea, 2005). We tried to minimize this by having...
the graduate TAs collect the assignments and by giving full credit for completing the assignments, no matter what the responses. An alternative explanation is that students taking this introductory biology course were primarily science and pre-professional majors who may be more invested in doing well in the course. It would be interesting to study students in a nonmajors introductory course to see whether undergraduates’ use of metacognitive-regulation skills depends on interest in the area of study.

CONCLUSION

To begin to characterize metacognition in undergraduates, we asked what metacognitive-regulation skills are evident in students taking an introductory biology course. Using exam preparation as a vehicle for studying metacognition, we found that approximately half of the students were able to monitor, evaluate, and plan the learning strategies they used to prepare for exams. Interestingly, nearly all of the students were willing to reflect and adjust their study plans, but many did not identify appropriate learning strategies, and many did not carry out their new plans. We conclude that postexam assignments encouraged students to engage in metacognitive regulation, but many of the students needed additional help with metacognitive knowledge before they could fully benefit from the metacognition prompts in these exercises.

We have proposed a continuum of metacognitive-regulation development that represents what we saw in introductory biology students (Figure 1). We have used this model to formulate new research hypotheses and to help instructors target metacognition more effectively in the classroom. By helping students enhance their metacognitive-regulation skills, we will help improve their performance in our courses and others. Importantly, we will also help them become self-regulated learners, which will serve them well beyond their time in college.

ACKNOWLEDGMENTS

The authors thank Dr. Steve Hines for his encouragement of this project, Dr. Tessa Andrews for helpful discussions on this study, and Drs. Peggy Brickman and Paula Lemons for valuable feedback on the manuscript. This work was funded by generous start-up funds from the University of Georgia (to J.D.S.) and a Washington State University College of Veterinary Medicine Education Research Grant (to J.D.S. and X.N.N.). The initial design of this study was developed as part of the Biology Scholars Program Research Residency (J.D.S.).

REFERENCES


Yin RK (2010). Qualitative Research from Start to Finish, New York: Guilford.


The Synthesis Map Is a Multidimensional Educational Tool That Provides Insight into Students’ Mental Models and Promotes Students’ Synthetic Knowledge Generation

Ryan A. Ortega* and Cynthia J. Brame†‡

*Department of Biomedical Engineering and †Department of Biological Sciences, Vanderbilt University, Nashville, TN 37235; ‡Vanderbilt Center for Teaching, Nashville, TN 37212

Submitted July 13, 2014; Revised February 10, 2015; Accepted February 12, 2015
Monitoring Editor: Mary Lee Ledbetter

Concept mapping was developed as a method of displaying and organizing hierarchical knowledge structures. Using the new, multidimensional presentation software Prezi, we have developed a new teaching technique designed to engage higher-level skills in the cognitive domain. This tool, synthesis mapping, is a natural evolution of concept mapping, which utilizes embedding to layer information within concepts. Prezi’s zooming user interface lets the author of the presentation use both depth as well as distance to show connections between data, ideas, and concepts. Students in the class Biology of Cancer created synthesis maps to illustrate their knowledge of tumorigenesis. Students used multiple organizational schemes to build their maps. We present an analysis of student work, placing special emphasis on organization within student maps and how the organization of knowledge structures in student maps can reveal strengths and weaknesses in student understanding or instruction. We also provide a discussion of best practices for instructors who would like to implement synthesis mapping in their classrooms.

INTRODUCTION

Experts within a field have rich, well-connected knowledge structures that allow them to rapidly retrieve information and see unexpected connections and patterns. On the other hand, novices to a field have not yet formed these well-connected knowledge structures and therefore may see information as a series of disconnected facts or as groups of facts within disconnected silos. According to the key work How People Learn, “to develop competence in an area of inquiry, students must: (a) have a deep foundation of factual knowledge, (b) understand facts and ideas in the context of a conceptual framework, and (c) organize knowledge in ways that facilitate retrieval and application” (National Research Council [NRC], 2000, p. 16). Thus, a focus of many college courses is students’ development of an organizing framework to help them establish a sustained understanding of the subject under study (Khodor et al., 2004).

When teaching an upper-level undergraduate course, Biology of Cancer, one of the authors (C.J.B.) had noted students’ difficulty with developing this type of organizational framework. We therefore created an assignment type, termed the synthesis map, as a tool to help students within this course develop an explicitly defined organizational framework to describe their understanding of carcinogenesis. Specifically, students were asked to construct a visual representation of their model of carcinogenesis using a cloud-based presentation tool that allows representation in multiple dimensions. The students’ model was developed over the course of a semester; students adapted their maps in response to their growing knowledge base and formative feedback from the instructor and peers.

The synthesis map assignment may be considered an evolution of the concept map. The concept map as a student learning tool was developed in the 1970s by J. D. Novak and
is based on the cognitive learning theory of David Ausubel, which stressed meaningful learning (Ausubel, 1962). The purpose of using concept maps as a student assignment in classroom learning is to have students explicitly establish a personalized, hierarchical organization of their understanding of a certain subject or concept. As a student learning strategy, an instructional tool, and a means for both formative and summative assessment of student understanding, this tool has improved science education (Novak, 1990). Successfully implemented concept map assignments inspire student metacognition by encouraging students to examine their own knowledge structures, and repeated use of concept maps can reveal how these knowledge structures change over time (Novak and Gowin, 1984). Concept mapping as an iterative exercise challenges students to organize their knowledge, analyze the validity and efficacy of that organization, and even produce novel or modified knowledge structures. It is a tool that engages the higher levels of Bloom’s taxonomy in the cognitive domain (Peresich et al., 1990; Anderson and Krathwohl, 2001). It has also been shown to stimulate positive learning behaviors in the affective domain, such as students’ ability and willingness to receive and respond to information (Krathwohl et al., 1964; Maas and Leauby, 2005). This increase in positive affective-domain behaviors not only encourages meaningful learning but also significantly reduces student anxiety about learning the biological sciences (Jegede et al., 1990).

As Internet use became more popular and accessible in the 1990s, software facilitating concept map creation became more readily available. One example of such software is the freely downloadable CmapTools software developed by the Institute for Human and Machine Cognition. This and other software have enhanced students’ ability to construct concept maps. In general, advances in technology allow for large advances to be made in how education tools are implemented, and in 2003, J. D. Novak commented that “we need to expand our efforts to disseminate the new educational ideas and tools” (Novak, 2003, p. 131). It is with this challenge in mind that we have developed the synthesis map, which we believe to be a natural evolution of the concept map.

Like a concept map, a synthesis map is a visual representation of a student’s understanding and knowledge structures, and it is capable of highlighting the connectivity (or lack thereof) between concepts in the student’s knowledge base. However, two-dimensional concept maps are limited in scale by how much information can fit on a single paper or screen. Synthesis maps work around this constraint by utilizing a third dimension of depth to embed layers of information within concepts. However, a synthesis map is not simply a set of layered concept maps. Ideally, a synthesis map can be used to show hierarchical knowledge structures in the same presentation as temporally arranged process diagrams and spatially represented physical structures pertinent to the subject of interest. When used appropriately, a synthesis map can simultaneously present a detailed model of a student’s understanding of several broad subjects, illustrate the connectivity of these subjects, and delve deeply into the minute details of each subject. We have developed the synthesis map exercise specifically to target the higher levels of Bloom’s taxonomy in the following ways: the assignment necessitates that the student analyze and categorize information, each student must construct and develop his/her own unique knowledge structure, and each student is required to evaluate and summarize research regarding his/her subject to provide evidentiary support for the model he/she has constructed.

In implementing this new type of learning and assessment tool, we used the currently available, free to use presentation software Prezi. Prezi is a cloud-based presentation tool that allows for seamless horizontal transitions, and it combines this utility with a zooming user interface (ZUI) that lets the author of the presentation use both depth and distance to show connections between data, ideas, and concepts (Conboy et al., 2012). Users can import images into the Prezi interface as well as create their own images and organizational structures and icons. It is possible to use Prezi to construct traditional concept maps without depth or to create a linear presentation similar to a PowerPoint presentation. Prezi, however, also allows the user to create multiple, hierarchical knowledge structures in the same presentation and provides the option of exploring these structures in a unique, nonlinear path (Rockinson-Szapkiw et al., 2011).

We have used the synthesis map assessment tool in the context of a biological sciences course covering the biology of cancer. This class is designed to provide an introduction to the underlying principles of cancer development emerging from the vast and growing collection of facts about this disease. The synthesis map has particular utility in biological science courses due to the high degree of connectivity present between biological concepts, the large range in physical scale in biological organisms (from single atoms to whole populations), and the great range of conceptual depth associated with biological knowledge (Smith et al., 2013). It is within the context of the range of scale and depth inherent in the biological sciences that Prezi’s ZUI can demonstrate its utility. Students in this class constructed synthesis maps as visual representations of their model of carcinogenesis. One of the challenges in understanding a complex process like carcinogenesis is fitting the different components into a coherent whole. By constructing visual representations of their model (which, by definition, changes in response to new knowledge), the students clarified and structured their growing understanding of carcinogenesis.

In this study, we examined the features of the synthesis maps students produced in this course, asking the following questions:

- What descriptive statistics do we observe for the maps produced in this course? Was any dimension of these descriptive statistics predictive of a more effective synthesis map?
- What organizational strategies did students use? Were these organizational strategies equally effective?
- Did elements of the synthesis map construction predict success on other, more traditional measures of student learning?

This study, therefore, falls within the “what is” category of the taxonomy of scholarship of teaching and learning studies seeking to describe constituent features of the synthesis map learning tool (Hutchings, 2000).

In this paper, we describe our observations of student organizational strategies used to create the synthesis maps and of how students’ organizational constructs can reveal the nuances of their understanding (or lack of understanding).
of the subject. We also scored the maps based on four criteria (organization between categories, organization within categories, accuracy, and completeness) and present those scores. Quantitative, correlative analysis shows that elements of students’ organization of the synthesis map interacts with other class metrics, suggesting a potential tool for revealing student understanding that is particularly important for success within a course.

One area in which more development is needed in educational research is recommendations for the practical application of concept mapping as an assessment method (Ruiz-Primo and Shavelson, 1996). To that end, we have included a set of best practices developed during retrospective analysis of the synthesis map assignment. It is our hope that this paper will provide a clear guide for other life sciences educators to utilize the synthesis map technique we have developed, leveraging current technology to enhance and evolve a well-established teaching tool.

METHODS

Recruitment of Students into the Study
The study was carried out in a biological sciences course at Vanderbilt University, a mid-sized, highly selective, private, research-intensive university in the southeastern United States. The course, which focused on the biology of cancer, is an elective course aimed at junior and senior undergraduates from a mix of science and engineering majors. Investigator C.J.B. was the instructor for the course. Students completed the synthesis map assignment as part of their normal course work. After completion of the course, students received two letters: one letter asking for permission to use their class data in the study and a second letter asking for permission to use images from their map as examples in the manuscript reporting study results. Copies of the letters are included in the Supplemental Material (pp. 1–2). The study was carried out under the approval of the Vanderbilt Institutional Review Board.

Out of 27 students, 24 agreed to participate in the study.

Introduction and Commencement of the Synthesis Map Assignment
The synthesis map project was introduced on the first day of class when reviewing the syllabus (see the Supplemental Material, pp. 3–8), which provided a brief description of the goal and format of the project. During the third week of class, the project was reviewed, and one class day was spent taking the first steps. Specifically, students self-assembled into groups of three and drafted their first synthesis map on paper based on the following prompt:

Work with two colleagues to draft your first synthesis map (20 Minutes).
First, compare your lists of things you know about cancer development. Discuss until you can come to a common list. If there is strong disagreement, then the lists for the three people don’t have to be identical.
Second, think about how you would represent these ideas visually. Sketch out at least one possible visual representation that incorporates all the things you know. It can include symbols, real images, labels, videos, etc. You can think in terms of scale; zooming in and out.

After the initial submission, each student received formative feedback from the instructor and from the two peers within his/her group. The groups remained constant throughout the semester. Students submitted their synthesis maps three more times during the semester, revising and expanding them as their knowledge grew. They received formative feedback from the instructor and two peers each time, with the feedback consisting of written comments identifying strengths and weaknesses in organization, clarity, accuracy, and completeness. Before the final submission, no grades were assigned. The midsemester submissions were important to the final grade, however, as each submission earned the student the “right” to 20% of the final grade; that is, a student who missed a midsemester submission could earn a maximum of 60/75 on the final synthesis map.

Grading
Student synthesis maps were initially graded by the course instructor (C.J.B.), using the grading rubric provided in the Supplemental Material (p. 9); the rubric was developed in collaboration with the students in the class and consisted of a total of 75 points. After completion of the course, the maps created by students consenting to be a part of this study were analyzed separately by both authors using an adapted form of the rubric used during the course. The maps were assessed based on the organization presented between major concepts associated with tumorigenesis, the organization within these major concepts, the accuracy of the information presented, and the completeness of the map. Each of these components was assigned a score from 0 to 10, with 10 being the best possible score. The total score for each map was the combined scores of these four components for a total of 40 possible points. The total number of slides the students used for each map were also counted, and the maximum level of embedding for each map was determined. Once individual scores were assigned, a consensus score for each category of each map was determined during analysis of each map.

Qualitative Analysis of Synthesis Maps
To explore the choices made by the students in constructing their synthesis maps, we analyzed the characteristics of their maps. We asked two questions about each map: “What organizational strategies did the student use?” and “What organizational strengths and weaknesses did we observe?” To address these questions, we used a modified grounded-theory approach (Strauss and Corbin, 1990). The
authors briefly reviewed all of the maps together, identifying potential themes. Each author then independently reviewed the maps for answers to these questions, categorized the responses, and then examined the categories to identify themes. We then compared our analyses, in most cases reconciling discrepancies to arrive at a single interpretation.

Statistical Analysis of Quantitative Data
The quantitative student data from the synthesis maps described above were collected, as were data regarding students’ exam grades, final paper grades, and final class grades. These quantitative components were analyzed for correlations by determining the Pearson’s product-moment correlation statistic and calculating the $p$ value associated with significance tests for these correlations. Correlation tests resulting in a Pearson’s statistic of $|1–0.7|$ were considered to have strong correlation, values of $[0.69–0.4]$ were considered to indicate moderate correlation, values of $[0.39–0.1]$ were considered to indicate weak correlation, and values less than $[0.1]$ were considered to indicate no correlation. Correlations analyzed by regression analysis were done with a simple linear regression model created using the ordinary least-squares method. The linear regressions are displayed in the standard form for a linear equation: $y = mx + b$. $p$ Values were calculated for each regression line. $p$ Values < 0.05 indicate that the predictor variable significantly explains some portion of the variation in the response variable. This portion is indicated by the $R^2$ value.

RESULTS

Qualitative Description: Synthesis Map Organizational Strategies

We examined the final synthesis maps for 24 students from a class of 27; the remaining three students either did not respond to requests for permission to examine the maps or had deleted the maps from their accounts. All of the synthesis maps represented seven major components of cancer development addressed in the course, although the depth and style of this representation varied: the progressive nature of carcinogenesis, the role of accumulating mutations and genetic rearrangements, proto-oncogenes, tumor suppressor genes, immortalization, interactions with noncancerous cells, and metastasis. All of the maps we examined were multilayered, with all students using the zooming feature of Prezi to embed information within larger concepts; the degree of layering varied from two layers to five layers. In addition, students uniformly created paths through their maps, allowing viewers to click through steps to track through the map. The lengths of these path varied, ranging from 56 to 141 “frames,” and students typically used arrows and guiding text to help viewers understand the organization. In some cases, the students created paths that encompassed all features of the map, while other students included additional features that could be explored independently of the planned pathway.

Students used three major organizational strategies when creating their maps: conceptual, spatial, and narrative:

- Students using conceptual organization identified particular features of cancer development, such as the components listed above or the hallmarks of cancer identified by Robert Weinberg (Hanahan and Weinberg, 2000), and used each of these as a node in their maps, describing each feature relatively independently of the others. In essence, maps using this organizational strategy resembled a collection of concept maps or a visual depiction of an outline. Figure 1 displays a screenshot of a synthesis map that relied primarily on conceptual organization.

- Students using spatial organization created structures that were explicitly meant to represent physical components associated with a given concept. For example, students using this organizational strategy often created images of normal and cancerous cells and embedded features that
We found that the maps offered insight into how students conceptualized relationships between big components of cancer development. For example, the map shown in Figure 1 uses different circles on the map to describe components of cancer development. The author does not make relationships between these components explicit; implicit links are apparent to a knowledgeable, careful viewer who notes example proteins that appear in two or more circles, but the author does not make these links explicit for the viewer. The author did, however, select many images from sources besides the textbook (∼40% of the total), including many that were not used in class materials. Furthermore, these images typically conveyed rich meaning within the map. Thus, the decision of this student to find images beyond those used in the class to illustrate a conceptualization of cancer suggests a well-developed knowledge structure, in spite of the lack of explicit connections between concepts.

We asked whether a particular organizational strategy (conceptual, spatial, or narrative) or combination of strategies was more effective than other combinations. The Venn diagrams in Figure 4 show that a majority of students used some form of conceptual organization in their synthesis maps. Synthesis maps utilizing only spatial or narrative arrangement alone or in combination were less popular options, but a breakdown of synthesis map scores and scores for organization indicates that all organizational schemes were conducive to successful synthesis map creation. All three primary strategies and their combinations appeared to allow effective map organization, although a purely conceptual organization—akin to a visual outline—appeared to be slightly less effective at conveying relationships between concepts.

We found that the maps offered insight into how students conceptualized relationships between big components of cancer development. For example, the map shown in Figure 1 uses different circles on the map to describe components of cancer development. The author does not make relationships between these components explicit; implicit links are apparent to a knowledgeable, careful viewer who notes example proteins that appear in two or more circles, but the author does not make these links explicit for the viewer. The author did, however, select many images from sources besides the textbook (∼40% of the total), including many that were not used in class materials. Furthermore, these images typically conveyed rich meaning within the map. Thus, the decision of this student to find images beyond those used in the class to illustrate a conceptualization of cancer suggests a well-developed knowledge structure, in spite of the lack of explicit connections between concepts.

The map shown in Figure 2 shows a highly integrated knowledge structure around carcinoma cell–stromal (“heterotypic”) interactions, angiogenesis, and metastasis, with the spatial organization of map facilitating a concise and highly visual integration of the topics. The map shown in Figure 3 displays another mechanism for relating components of cancer, using questions or short statements, often narrated via the Reveal function in Prezi, to draw explicit relationships between topics (see frames 35, 66, 82–83, 88, and 93).

Figure 2. (A) Screenshot showing an overview of a synthesis map that used primarily spatial organization; you can explore the map and follow the author’s suggested path by following this link: http://prezi.com/tp6jcnmlh3a/synthesis-map-complete-version/?utm_campaign=share&utm_medium=copy. (B) A closer view of frame 9 of the map, which illustrates finer details of the spatial organization of information. The student has created a representation of a cell and its nucleus. (C) Frame 31 of the same synthesis map shows a representative image of how some students incorporated images of transmembrane receptors at the cell surface into their spatial model of a tumor cell. Sources for images used within the map are provided in the Supplemental Material.

Figure 3. A screenshot of a synthesis map that relied heavily on narrative organization.
Quantitative Analysis

Effective Synthesis Map Features. Each synthesis map was scored for organization (both between concepts and within concepts), accuracy, and completeness using the rubric described in the Supplemental Material. Overall, student maps were very accurate, complete, and well organized (Table 1). The average score on the project was 84%, with one student earning a perfect score. The distribution of student scores shown in a histogram (Figure 5) visually approximates a normal distribution centered around a “B+” or “A−” grade.

Multiple students achieved perfect scores on individual metrics. Of the quantitatively assessed metrics, map accuracy had the highest average score and the smallest deviation. Although some students only minimally utilized the embedding feature of the Prezi software, every student created embedded information in his/her map to some extent using the ZUI, with most students implementing more than two levels of depth to indicate that a certain piece of information or a concept belonged in a hierarchical subset of the one above it. In other words, the student appeared to get lost in the details of a particular node, then jump out, begin another node, and repeat the process.

The maps also provided insight into students’ incorporation of evidence into their personally synthesized knowledge structures. For example, the map shown in Figure 3 repeatedly incorporates evidence into the section on cancer progression (see, e.g., frames 13, 14, 16–22). None of the maps examined, however, incorporated significant evidence into the sections on proto-oncogenes and tumor suppressor genes, suggesting a possible deficit in instructional materials.

and 92–94 for examples). The connections that are missing or unclear can also be informative. For example, the map shown in Figure 3 includes p53 and Rb as examples of tumor suppressor proteins but does not link inactivation of these proteins to cell immortalization through escape of senescence. This absence may reveal mental “siloing” of information rather than a fully integrated understanding of these related topics.

Other examples revealed larger difficulties with organizing information. One map we investigated exhibited several completely unconnected nodes within the map. Within several of these nodes, there were large blocks of very detailed text and multiple images supporting the text. Each node, however, would be incomplete and unconnected from other nodes. The overall appearance of the map was disorganized and scattered. Taken together, these observations suggested the student was having difficulty fitting detailed information into a large, integrated framework. In other words, the student appeared to get lost in the details of a particular node, then jump out, begin another node, and repeat the process.

The connections that are missing or unclear can also be informative. For example, the map shown in Figure 3 includes p53 and Rb as examples of tumor suppressor proteins but does not link inactivation of these proteins to cell immortalization through escape of senescence. This absence may reveal mental “siloing” of information rather than a fully integrated understanding of these related topics.

Other examples revealed larger difficulties with organizing information. One map we investigated exhibited several completely unconnected nodes within the map. Within several of these nodes, there were large blocks of very detailed text and multiple images supporting the text. Each node, however, would be incomplete and unconnected from other nodes. The overall appearance of the map was disorganized and scattered. Taken together, these observations suggested the student was having difficulty fitting detailed information into a large, integrated framework. In other words, the student appeared to get lost in the details of a particular node, then jump out, begin another node, and repeat the process.

The maps also provided insight into students’ incorporation of evidence into their personally synthesized knowledge structures. For example, the map shown in Figure 3 repeatedly incorporates evidence into the section on cancer progression (see, e.g., frames 13, 14, 16–22). None of the maps examined, however, incorporated significant evidence into the sections on proto-oncogenes and tumor suppressor genes, suggesting a possible deficit in instructional materials.

Quantitative Analysis

Effective Synthesis Map Features. Each synthesis map was scored for organization (both between concepts and within concepts), accuracy, and completeness using the rubric described in the Supplemental Material. Overall, student maps were very accurate, complete, and well organized (Table 1). The average score on the project was 84%, with one student earning a perfect score. The distribution of student scores shown in a histogram (Figure 5) visually approximates a normal distribution centered around a “B+” or “A−” grade. Multiple students achieved perfect scores on individual metrics. Of the quantitatively assessed metrics, map accuracy had the highest average score and the smallest deviation.

Although some students only minimally utilized the embedding feature of the Prezi software, every student created embedded information in his/her map to some extent using the ZUI, with most students implementing more than two levels of depth to indicate that a certain piece of information or a concept belonged in a hierarchical subset of the one above it. The largest range in any given metric was in the number of slides used to fabricate each student’s map.

The set of scores given in each metric described above were compared with one another to determine whether there were correlations between different metrics making up the synthesis map assignment. The Pearson product-moment correlation statistics were calculated to determine correlation between the different data sets (Table 2), and correlations were checked for significance by calculating the p value for each correlation (Table 3). For the Pearson statistic,
The correlations between different components of the synthesis map reveal important information about the learning tool, not only with the components that are significantly correlated but also with the components that are not significantly correlated. For example, it is expected that the two organization categories, as well as completeness and accuracy would correlate moderately or strongly with the total score, a $p$ value $< 0.05$ means that we can reject the hypothesis that the correlation is due to random sampling, indicating that the correlation is significant.

![Venn diagrams](image)

**Figure 4.** Venn diagrams describing the distribution of organizational strategies used on students’ synthesis maps and matching organizational strategies to average values for map scores. (A) The number of students utilizing each organizational strategy alone or in combination. (B) The average score on the synthesis map (out of 40) for each organizational scheme ($\pm \sigma$). (C) The average score for organization between major concepts on the synthesis map (out of 10) for each organizational scheme. (D) The average score for organization within individual concepts on the synthesis map (out of 10) for each organizational scheme.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organization between major concepts</td>
<td>8.16 out of 10</td>
<td>0.30</td>
<td>4–10</td>
</tr>
<tr>
<td>Organization within major concepts</td>
<td>8 out of 10</td>
<td>0.29</td>
<td>4–10</td>
</tr>
<tr>
<td>Completeness</td>
<td>8.33 out of 10</td>
<td>0.23</td>
<td>6–10</td>
</tr>
<tr>
<td>Accuracy</td>
<td>9.21 out of 10</td>
<td>0.19</td>
<td>7–10</td>
</tr>
<tr>
<td>Total score</td>
<td>33.7 out of 40</td>
<td>3.86</td>
<td>21–40</td>
</tr>
<tr>
<td>Maximum levels of embedding</td>
<td>3.33</td>
<td>0.16</td>
<td>2–5</td>
</tr>
<tr>
<td>Number of slides in presentation</td>
<td>98</td>
<td>4.8</td>
<td>56–141</td>
</tr>
</tbody>
</table>

**Table 1.** Statistics for grade values given to student synthesis maps covering organization, completeness, and accuracy, as well as for synthesis map total score and descriptive metrics: maximum levels of embedding and number of slides.

![Histogram](image)

**Figure 5.** Histogram of student scores on synthesis map assignment (out of 40 possible points).
because the sum of the score from these components makes up the total score. However, the maximum levels of embedding used in the map is significantly, moderately correlated with the total score \( (p = 4.3 \times 10^{-4}) \), indicating that, even though this component did not numerically contribute to the total grade, it did contribute in other, less obvious ways. Antithetically, the total number of slides used in each presentation did not significantly correlate to the total score. In fact, the only significant correlation for the number of slides used is a moderate correlation with completeness, indicating that, while the number of slides does not have a strong effect on organization or accuracy, it does have an effect on the completeness or perceived completeness of the map. Going further, it implies that there is a minimum number of slides that must be used to incorporate all of the necessary information for this specific assignment.

The synthesis map assignment is designed to provide insight into students’ knowledge structures, so the organization of the maps is of particular interest. The scores for organization between major concepts do not significantly correlate with the scores for organization within individual topics covered by the maps, but the \( p \) value \( (p = 0.071) \) is close enough to significant to indicate that a larger sample size may be needed to show true significance. However, the organizational strategies and goals for these two categories are not necessarily similar. This is evidenced by the fact that one of the measurable metrics of organization, the amount of embedding used in the map, correlates significantly with the student scores for organization within concepts but does not correlate with organization between concepts. By its very nature, this capability of the Prezi software lends itself to increasing the amount of fine detail and showing the underlying organizational structure within the structure in which the data or concepts are embedded. It is also interesting to note that the level of embedding has a significant moderate correlation with map accuracy, indicating that one component of accuracy may be the way that the fine detail of the knowledge structure is arranged. Certainly, a clearly, deeply organized map presents a greater appearance of accuracy than a cluttered or shallow map.

**Correlation with Other Class Metrics.** To determine whether any element of the synthesis map could provide particular value for promoting student success, we analyzed the map scores and the components of the map scores against other metrics of class success: combined exam scores, final paper scores, and class grade without the synthesis map included. Table 4 shows data extracted from correlation tests between components of the synthesis map and other class metrics. Notably, we observed a significant, moderate correlation between students’ ability to illustrate organization between concepts and their performance on course exams and in the final course grade (Table 4, rows 1 and 4).

Linear regressions were performed following correlative analysis of student data to determine whether certain components of the synthesis map could act as linear predictor variables for other measurements of success in the course. The organization between major concepts in the map is of particular interest, because it has significant correlation with class score and total exam score. Figure 6 shows linear regressions with organization between major concepts on the synthesis map as the predictor variable for class score and total exam score. The variation in scores for organization between concepts on the synthesis map can account for

![Table 2. Pearson’s product-moment correlation statistic \( (r) \) for different components of the synthesis map assignment](image)

<table>
<thead>
<tr>
<th></th>
<th>Organization between concepts</th>
<th>Organization within concepts</th>
<th>Completeness</th>
<th>Accuracy</th>
<th>Levels of embedding</th>
<th>Number of slides</th>
<th>Total score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organization between concepts</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organization within concepts</td>
<td>0.58</td>
<td>0.43</td>
<td>0.61</td>
<td>0.51</td>
<td>0.30</td>
<td>0.55</td>
<td>0.21</td>
</tr>
<tr>
<td>Completeness</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.54</td>
<td>0.69</td>
<td>0.37</td>
<td>0.51</td>
<td>0.30</td>
<td>0.55</td>
<td>0.21</td>
</tr>
<tr>
<td>Levels of embedding</td>
<td>0.26</td>
<td>0.59</td>
<td>0.37</td>
<td>0.51</td>
<td>0.30</td>
<td>0.55</td>
<td>0.21</td>
</tr>
<tr>
<td>Number of slides</td>
<td>0.06</td>
<td>0.08</td>
<td>0.54</td>
<td>0.03</td>
<td>0.30</td>
<td>0.55</td>
<td>0.21</td>
</tr>
<tr>
<td>Total score</td>
<td>0.79</td>
<td>0.88</td>
<td>0.61</td>
<td>0.80</td>
<td>0.55</td>
<td>0.21</td>
<td>1</td>
</tr>
</tbody>
</table>

**Significant \( p \) values are indicated in bold.**

![Table 3. \( p \) Value \( (p) \) for the correlation statistics for different components of the synthesis map assignment](image)

<table>
<thead>
<tr>
<th></th>
<th>Organization between concepts</th>
<th>Organization within concepts</th>
<th>Completeness</th>
<th>Accuracy</th>
<th>Levels of embedding</th>
<th>Number of slides</th>
<th>Total score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organization between concepts</td>
<td>0.071</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organization within concepts</td>
<td>0.89</td>
<td>0.21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Completeness</td>
<td>0.083</td>
<td>0.001</td>
<td>0.26</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.15</td>
<td>0.005</td>
<td>0.16</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Levels of embedding</td>
<td>0.67</td>
<td>0.80</td>
<td>0.002</td>
<td>0.93</td>
<td>0.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of slides</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total score</td>
<td>0.008</td>
<td>1.5 \times 10^{-6}</td>
<td>0.01</td>
<td>3 \times 10^{-4}</td>
<td>4.3 \times 10^{-4}</td>
<td>0.17</td>
<td></td>
</tr>
</tbody>
</table>

**Significant \( p \) values are indicated in bold.**
~33% of the variation in combined exam score and 25% of the variation in class grade. This illustrates the power of having students examine and make explicit their knowledge structures. The score for organization between concepts makes up a small portion of the final grade for the synthesis map, which is itself only 14% of the course grade. Yet this metric can account for a large portion of the variation in class grade. These results suggest that formative feedback on this element of the synthesis map may have potential larger effects for student performance in a course.

**DISCUSSION**

The **Organization of Students’ Synthesis Maps Provides Insight into Their Knowledge Structures and Engages Meaningful Student Learning**

The way students organize their maps with respect to relationships between the major concepts of tumorigenesis represents their broad understanding of these concepts and their large-scale mental models interconnecting these subjects. The way students organize the data and evidence presented within each major subject or category provides insight into the finer structure of their understanding, as well as how they evaluate and incorporate real data into their personally synthesized knowledge structure. It is the combination of these two organizational metrics that reveals the hierarchical nature of the students’ understanding and shows how they transition from gross- to fine-scale structuring of knowledge regarding the chosen topic.

As with concept maps, the organizational structure of a synthesis map is most important for revealing student misconceptions and problems with instruction. The maps explicitly visualize the students’ strengths and weakness at different organizational scales. For example, some maps had strong large-scale organization between major concepts but had little data, detail, or evidence. This indicated that these students may have had a strong, gestalt view of tumorigenesis but either had not internalized information associated with the fine detail of these concepts or lacked the deep cognitive organizational structure for these concepts. Other maps had a large amount of fine detail within an individual concept but little connectivity or an incomplete final structure. Some maps existed between these extremes, indicating where students’ intermediate understanding or mental organizations might lie. Individual student’s misunderstanding can be diagnosed by using the synthesis map as a formative assessment tool and looking for errors in knowledge structure, errors in content, or omission of content (McClure et al., 1999). Problems with instruction will likely lead to multiple maps displaying the same or similar error(s). Once discovered, errors or omissions in instruction can be corrected in class and/or the syllabus can be adjusted for future classes.

By examining and testing student knowledge structures with synthesis mapping, it is possible to engage students in a different way compared with purely objective assessment methods such as multiple choice and short answer. It is known that concept map scores do not correlate well with multiple-choice scores, indicating that mapping exercises assess different skills, such as higher-level skills in the cognitive domain of Bloom’s taxonomy (Morse and Jutras, 2008). This different method of engagement with course material can have a positive impact on students for whom lecture and text is not the optimum learning path. The relational knowledge created with a synthesis map is something that learners must construct for themselves, and the act of creating relationships between existing knowledge and new knowledge creates meaningful learning (Novak, 2003).
Our observation that students’ ability to illustrate organization between major concepts can predict their success in other measures (i.e., test scores and final grade) corresponds to Bransford and colleagues’ identification of conceptual frameworks as key to learning (NRC, 2000). It may also suggest that helping students develop a broad, well-connected conceptual framework early in the course may improve their overall learning. By explicitly helping students develop such a broad framework before moving to more nuanced, data-driven details, instructors may be able to accelerate their students’ progression toward expert-like knowledge structures. This observation may be particularly relevant in cell and molecular biology courses like the biology of cancer course used as an example here, wherein a highly inductive approach to research may lead to courses that focus on concepts emerging from examples.

**Suggested Good Practices for Implementing the Synthesis Map**

Teaching requires simultaneous allocation of student and instructor resources toward the pursuit of multiple goals. It must be asked of each new teaching technique: Will it be worth the effort? We have included some suggested best practices for instructors implementing the synthesis map that will increase the efficacy of this teaching tool and reduce wasted effort:

**Introduce the Assignment Early in the Semester.** One concern with concept mapping as a teaching tool is that it takes a significant amount of time to teach and implement concept mapping in the classroom, especially with students with no practical exposure to the technique (Maas and Leauby, 2005). Synthesis mapping is inherently more complex than concept mapping, so it requires an even greater time commitment and more practice/procedural instruction.

**State Explicitly the Intended Audience.** Synthesis maps could be used for a variety of purposes: to share only with the instructor to reveal knowledge structures, to share with peers within a class to help with studying or to teach individual topics related to larger class themes, to share on the Web for general education purposes, to use in teaching projects with younger students. Each of these possibilities represents a different audience that can be expected to bring different starting knowledge. Being explicit with students about the intended audience will help them consider the background they need to provide and the level of detail that will be appropriate.

**Describe Potential Organizational Strategies.** To allow students freedom to develop their own visual models, we intentionally gave very sparse instructions about what the synthesis maps should look like. While this approach was effective for most of the students in the course, it was more comfortable for some students than for others, some of whom vacillated between organizational strategies. Describing potential organizational strategies that others have used, however, may allow some students to intentionally choose the approach that best fits their mental model.

**Encourage Students to Make a Sparse Map Early in the Process and Have Them Revise and Iterate.** The most successful students chose an organizational strategy and used it consistently throughout their map. Some students started their map from scratch late in the semester, disregarding their earlier work. These examples generally had a more sophisticated and coherent organization but often lacked completeness. We hypothesize that this was caused by fatigue associated with attempting to complete the entire assignment from scratch late in the semester.

**Have the Students Work in Groups to Provide Peer Feedback and Affective Support.** Students working on concept maps in teams have been shown to be more successful on the assignment (Morse and Jutras, 2008). With the synthesis map assignment, we also saw instances of convergent design in many groups. In these groups, students can learn from one another and receive regular peer feedback. We hypothesize that this feedback provides positive affective support to students and increases problem-solving performance and success on the assignment. One negative effect of small groups that must be acknowledged is the possibility that convergent design could potentially mask students’ personal understanding or force them to adopt another student’s organization.

**Allow Multiple Instances for Feedback for the Students, Which Can Come from Other Members of Their Small Groups and from More Formalized Instructor Feedback during the Semester.** The fundamental process of concept mapping is not necessarily iterative; formative assessment using concept maps can take just minutes of class time. Owing to the increased complexity inherent in creating a synthesis map, the process should go through several iterations of map building, followed by feedback and modification (Allen and Tanner, 2003). By providing regular, ungraded feedback, instructors can use the synthesis map as another method of teaching throughout the course and as a means to help students address their misconceptions and to assess student learning at the end of the course. Furthermore, by regularly examining student synthesis maps, an instructor can achieve increased awareness of potential errors in instruction and can correct these errors as needed.

**Instruct the Students to Make Use of the Embedding Feature and Be Able to Show Examples of Effective Embedding.** Our results indicate that effective use of the ZUI to embed information improves the organization of students’ synthesis maps. This is particularly important, because the organization of hierarchical knowledge maps is more important to student success on the assignment than the actual information being presented (Morse and Jutras, 2008). However, the embedding function may not be intuitive to all students. That is why it is helpful if instructors provide examples of what they consider to be effective embedding when introducing the students to this feature. It is important to note that this is not the same as an instructor creating a “master map” or answer key against which student synthesis maps can be graded. While the use of a master map may be helpful for using concept maps as a summative assessment tool, the added level of organizational freedom offered by the Prezi software allows extreme variation in map design. While small, student-produced concept maps may converge on what can be considered the “correct” organization, this convergence was not seen with synthesis maps (McClure et al., 1999).
Encourage Students to Create Their Own Representative Structures and Diagrams in the Map Rather Than Relying on Premade Structures and Published Images. Students who created their own visual structures as opposed to using images from texts or journal articles were very successful. For example, some students used the available software to create their own representations of pathway diagrams, including proteins in the pathways, which were seamlessly integrated into other structures in the presentation. This allowed the students to completely realize their personal mental organization of these concepts.

Recognize the Value of the Map by Assigning It a Significant Number of Points. Synthesis maps that represent a large fraction of the course content and that are iteratively revised require a significant time investment from the instructor and from students. To make the assignment worth the time invested, it is important to allot a substantial number of points to its completion. As with any assignment, the actual point value and time allotment must depend on the place that the synthesis map occupies in the overall course plan. For example, if an exam is worth 25% of the points within a course and should require studying seven out-of-class hours per week for 4 wk, then it would be reasonable to do a similar calculation for a synthesis map covering the same scope. Importantly, evaluating synthesis maps can be similar to evaluating student writing, with the same potential for time consumption. It is therefore important to set guidelines to help keep your grading efficiency in line with your expectations for the project. Education blogs provide multiple suggestions to enhance efficiency, one example being http://mssphillips.wordpress.com/2014/03/10/giving-better-feedback-google-form-rubrics-and-autocrat.

CONCLUSION

Teachers are challenged to apply technological advances to the instruction and assessment given to their students. Synthesis mapping uses a novel, multidimensional presentation tool to allow students to create a detailed map of their hierarchical knowledge structures. By leveraging the ZUI to embed knowledge and concepts within the map, students can create maps of greater complexity than what is allowed for using traditional concept-mapping tools and strategies. This increase in scope and complexity helps students engage higher-level cognitive skills. Synthesis mapping is an explicitly constructivist tool, in that it directly models the constructivist process. Like the concept map, it can be used to promote, and even force, students to take a metacognitive approach to understanding their own knowledge structures. It also challenges students in ways that other course-assessment methods do not. Knowledge mapping is a very effective formative assessment tool. When used successfully, synthesis mapping can be used as formative assessment to reveal students’ strengths, misconceptions, and organizational schema, and as a summative assessment to test students’ understanding of course material and their ability to use and evaluate that material.

ACKNOWLEDGMENTS

The authors thank the students who agreed to participate in this study. In particular, we thank John Cao, Saad Rehman, and Emily Summerbell for allowing us to publish images from their synthesis maps and provide links to their maps in this article.

REFERENCES


Research and theory development in cognitive psychology and science education research remain largely isolated. Biology education researchers have documented persistent scientifically inaccurate ideas, often termed misconceptions, among biology students across biological domains. In parallel, cognitive and developmental psychologists have described intuitive conceptual systems—teleological, essentialist, and anthropocentric thinking—that humans use to reason about biology. We hypothesize that seemingly unrelated biological misconceptions may have common origins in these intuitive ways of knowing, termed cognitive construals. We presented 137 undergraduate biology majors and nonmajors with six biological misconceptions. They indicated their agreement with each statement, and explained their rationale for their response. Results indicate frequent agreement with misconceptions, and frequent use of construal-based reasoning among both biology majors and nonmajors in their written explanations. Moreover, results also show associations between specific construals and the misconceptions hypothesized to arise from those construals. Strikingly, such associations were stronger among biology majors than nonmajors. These results demonstrate important linkages between intuitive ways of thinking and misconceptions in discipline-based reasoning, and raise questions about the origins, persistence, and generality of relations between intuitive reasoning and biological misconceptions.
biology educators. Finally, we present hypotheses about potential interactions between these two seemingly unrelated arenas—intuitive ways of knowing from psychology and misconceptions observed in formal biology education.

Intuitive Ways of Thinking and Cognitive Construals
Three decades of cognitive science research have demonstrated that humans naturally, intuitively, and effortlessly reason about biological entities, processes, and phenomena in predictable ways (e.g., Carey, 1985; Berlin, 1992; Inagaki and Hatano, 2002, 2006; Coley et al., 2002; Atran and Medin, 2008). We have dubbed these regularities cognitive construals (Coley and Tanner, 2012). A cognitive construal is an informal, intuitive way of thinking about the world. It might be a set of assumptions, a type of explanation, or a predisposition to a particular type of reasoning. Three such cognitive construals—teleological thinking, essentialist thinking, and anthropocentric thinking—are common themes spanning research on intuitive biological thought. We hypothesize that they may also have particular relevance to understanding challenges and misconceptions commonly encountered when students are learning life science concepts. Each is reviewed briefly in the following sections.

Teleological Thinking. Cognitive psychologists have shown that our minds are biased toward causal explanations (e.g., Gopnik, 2000; Sloman, 2005; Kahneman, 2012). We are quick to generate causal stories for events, from an uptick in the stock market to an above-average yield of tomatoes. Explaining an event by reference to the outcome or consequences of that event, rather than an antecedent of the event, is known as teleological thinking (Keil, 2006; Talanquer, 2007, 2013). In other words, teleological thinking is causal reasoning based on the assumption of a goal, purpose, or function. Kelemen (1999) argues that teleological thinking is a central component of adults’ everyday thought. We make the teleological assumption that people’s actions are directed toward certain goals and presume that human artifacts, such as chairs and coats, are designed by their creators to fulfill some intended purpose. As Kelemen emphasizes, teleological thinking provides an important component of adults’ intuitive interpretations of why events occur or why objects have the properties they do.

Essentialist Thinking. Essentialist thinking is the tendency to believe that a core underlying property or feature of a biological structure, species, or system determines its overt features and identity. This cognitive construal is an assumption that people make about concepts. For cognitive scientists, concepts are mental representations of categories, along with related knowledge (Murphy, 2002). Essentialist thinking captures the idea that in addition to summarizing knowledge, concepts also involve a possibly implicit assumption that there is some unobservable essential property (an “underlying reality” or “true nature”) common to members of a category that conveys identity and causes observable similarities among category members (Medin and Ortony, 1989; Ahn et al., 2001; Gelman, 2003).

One consequence of essentialist thinking is the belief that members of a category are relatively uniform with respect to shared properties—a shared essence should give rise to similar properties in all category members (e.g., Shutulman and Shulz, 2008), leading us to discard variability among category members as noise. An additional consequence of essentialist thinking is that superficial transformations (e.g., changes in appearance) should not affect category membership, which is ultimately based on the presence or lack of an essential property, rather than superficial features (Keil, 1989; Rips, 1989). A third is that category membership conveys innate potential; because of an underlying essence, category members not only share properties but also the propensity to develop certain characteristics over time (e.g., Gelman and Wellman, 1991; Solomon et al., 1996). In sum, essentialist thinking yields assumptions about uniformity and predictability of category members that reduce the complexity of incoming information to manageable levels.

Anthropocentric Thinking. Anthropocentric thinking involves distorting the place of human beings in the natural world. This can result in 1) the tendency to see humans as unique and biologically discontinuous with the rest of the animal world and 2) the tendency to reason about other organisms by analogy to humans. The first component involves the way in which human beings are incorporated into the intuitive taxonomy of living things. According to geneticists, humans are African great apes; we share a common ancestor who lived c. 5–8 million yr ago with our closest living relatives: chimpanzees. However, intuitive biological taxonomies—particularly those found in industrialized Western societies—tend to see humans as essentially separate and discontinuous from other species (e.g., Coley, 2007). The second component of anthropocentric thinking is the tendency to reason about unfamiliar biological species or processes by analogy to humans. Analogical reasoning—trying to understand an unfamiliar idea or situation by comparing it with something familiar—is a common strategy used across many domains of learning (Gentner and Smith, 2013). Intuitively, human beings are a familiar and accessible biological kind and are therefore a very tempting source of knowledge that is often misapplied to nonhuman living things. This can lead to both overattribution of human characteristics to similar organisms (e.g., Inagaki and Hatano, 1991) and underattribution of biological universals to dissimilar organisms (e.g., Carey, 1985).

Other Work on Intuitive Explanatory Frameworks. Our delineation of intuitive biological thought in terms of these foundational construals is very similar to the approach taken by Talanquer. Specifically, Talanquer has characterized in detail an intuitive explanatory framework in the domain of commonsense chemistry (Talanquer, 2006). This includes both “empirical assumptions” (including teleological and essentialist thinking, as applied to understanding chemical substances and processes) and “reasoning heuristics.” Indeed, Talanquer has documented the presence of teleological thinking in particular in both common chemistry textbooks (Talanquer, 2007) and explanations about specific types of chemical reactivity (Talanquer, 2010, 2013). Likewise, Taber and Garcia-Franco (2010) have identified a number of spontaneous, intuitive ways of thinking about chemistry among English secondary students. We see this as a fruitful approach to understanding systematic misconceptions and
take the inquiry a step further by documenting specific linkages between construals and misconceptions in the domain of biology.

In a similar vein, Evans and colleagues (2012) emphasize the importance of these cognitive construals as intuitive explanatory frameworks that contribute to the emergence of both understanding of and misconceptions about key concepts of evolution. Like Evans and colleagues, we acknowledge the importance of linking emergent scientific understanding with the development of intuitive conceptual frameworks. However, as detailed in the following sections, we see this linkage as relevant to a wider array of concepts across the life sciences.

**Misconceptions Observed in Formal Biology Education**

While psychology researchers have developed theoretical frameworks for understanding intuitive reasoning, science education researchers have independently documented a variety of scientifically inaccurate biological ideas held by biological novices. These misconceptions (also referred to as alternative conceptions, naïve conceptions, or preconceptions) are characterized as differing from a lack of knowledge in that they appear to be tenacious and retained in the presence of formal instruction unless explicitly addressed (Wandersee et al., 1994). Misconceptions appear to cut across boundaries such as age, ability, gender, and culture; moreover, instructors themselves have been shown to hold some of the same misconceptions as students (e.g., Arnaudin and Mintzes, 1985; Nehm and Schonfeld, 2007). Some misconceptions have been hypothesized to originate in or be exacerbated by formal education, such as biology instructors’ use of the term “adapt” in teaching evolution, leading students to conceptualize individual organisms rather than populations, changing over time. However, biology educators at all levels tend to address misconceptions individually, without a more systematic approach to teaching and supporting students in grappling with these ideas.

To the extent that misconceptions are seen as interrelated in the biology education literature, they are usually organized in terms of subject matter, and instructors tend to diagnose misconceptions using subject-based concept inventories, in areas such as evolution (e.g., Anderson et al., 2002), genetics (e.g., Smith et al., 2008), molecular biology (e.g., Shi et al., 2010), animal development (e.g., Knight and Wood, 2005), and microbiology (e.g., Marbach-Ad et al., 2009), to name a few. Some science education investigators have moved beyond analysis of individual misconceptions and have begun investigating patterns observed in students’ explanations of scientific phenomena. Several scholars have explored the nature of students’ explanations, connecting characteristics of students’ explanations to models of causal reasoning and to models of scientific explanation drawn from philosophy of science writings (e.g., Grotzer, 2003; Braaten and Windschitl, 2011). Others have described patterns in students’ intuitive reasoning, which includes misconceptions, and have organized these findings into learning progressions that characterize student reasoning related to carbon and water cycling (Mohan et al., 2009; Hartley et al., 2011; Gunckel et al., 2012).

The study of the nature and defining characteristics of misconceptions is an ongoing area of research. Some argue that misconceptions represent fragmentary, isolated “phenomenological primitives” (diSessa, 1988, 1993; Maskewicz and Lineback, 2013). Others argue that misconceptions grow out of coherent intuitive conceptual frameworks used to understand, explain, and predict the world (e.g., Vosniadou, 1994, 2002; Taber and Garcia-Franco, 2010; Coley and Tanner, 2012). All agree that the presence of misconceptions does not indicate deficits but rather a mind actively engaged with the world trying to construct explanations for complex phenomena. As such, misconceptions should not be regarded as simply wrong ideas to be fixed but rather as common ways of thinking that can be important starting points for teaching and learning (e.g., Leonard et al., 2014). In fact, misconceptions are likely key in driving conceptual change in educational settings and engaging learners in recognizing their existing conceptions about how the world works, examining these ideas in the context of scientific evidence, and then changing and restructuring their conceptions toward more evidence-based, scientifically accurate ideas held by experts in a scientific field (Posner et al., 1982).

**Hypothesized Relations between Cognitive Construals and Biological Misconceptions**

Thus, cognitive science research has suggested that intuitive biological thought can be characterized by at least three distinct cognitive construals: teleological thinking, essentialist thinking, and anthropocentric thinking. Science education research has revealed a plethora of misconceptions that have generally been considered independent of one another. Unfortunately, scholarship in these disciplines remains largely isolated, and genuine disciplinary border crossing is rare. To address the dearth of interdisciplinary thinking, we have proposed a new theoretical framework that derives from considering the science education research literature in light of the cognitive psychology literature (Coley and Tanner, 2012). Specifically, we hypothesize that the hallmarks of intuitive biological thought most commonly studied in children by cognitive psychologists persist in the conceptual systems of undergraduate students. We hypothesize that what emerges when these students encounter university-level biological science are systematic clusters of biological misconceptions that transcend standard biological subjects like evolution and cellular respiration but share an underlying cognitive and conceptual basis. If this is so, seemingly disparate biological misconceptions across diverse subject areas may have common origins in underlying cognitive construals (see Table 1).

There is some empirical support for this framework; a small but important body of research within cognitive science has linked misconceptions about evolution to construal-based reasoning. For example, Evans (2008) has argued eloquently for the importance of intuitive construals in understanding the difficulties involved in learning and teaching evolutionary theory (see also Coley and Muratore, 2012; Gelman and Rhodes, 2012; Kelemen, 2012; Shultsman and Calabi, 2012). Shultsman and Schulz (2008) have shown that undergraduates who perceive animal categories as homogeneous—one aspect of essentialist thinking—have a less sophisticated and
a previous study has shown that teleological narratives are less likely than anthropocentric narratives to be variable. Similarly, Kelemen and Rossett (2009) have shown that higher levels of teleological thinking are associated with poorer performance on the Conceptual Inventory of Natural Selection (Anderson et al., 2002). In addition, Legare et al. (2013) have shown that anthropocentric narratives are less likely than scientifically accurate narratives to facilitate accurate evolutionary understanding in elementary school children. These studies show some linkages between cognitive construals and understanding of evolution but have not examined relations between cognitive construals and understanding biological science more generally.

On the other hand, biology education researchers have documented the existence of teleological and anthropocentric thinking among high school and university students (e.g., Tamir and Zohar, 1991; Friedler et al., 1993; Zohar and Ginossar, 1998; Nehm and Ridgeway, 2011). Likewise, chemistry education researchers have systematically detailed intuitive frameworks for organizing knowledge about chemistry (e.g., Talanquer, 2006; Taber and Garcia-Franco, 2010) but have not systematically explored linkages between cognitive construals and systems of misconceptions. In the following sections, we propose to build on this foundation by investigating the relations between a wider range of biological misconceptions and the three cognitive construals that we hypothesize may give rise to these misconceptions.

**Hypothesized Teleological Misconceptions.** Teleological thinking is a widespread cognitive construal that is useful in helping us make sense of many aspects of the world around us. However, this natural form of explanation is often extended inappropriately in the domain of biology. Students at all levels commonly explain biological structures and processes by reference to their supposed purpose, goal, or function. The first section of Table 1 lists examples of misconceptions documented by multiple research groups. These span traditional subject areas in the life sciences, including plant respiration, evolution, and cellular development, but we hypothesize that they may all stem from teleological thinking. What these misconceptions share is a sense of forward-looking, goal-directed, outcome-driven causality that can obscure the underlying biological processes involved. Thus, we hypothesize that teleological thinking may underlie a variety of seemingly unrelated biological misconceptions and may thereby play a role in hindering students’ transitions from novices to expert thinkers in biology.

**Hypothesized Essentialist Misconceptions.** Essentialist thinking, which includes the assumption that underlying essential properties cause external features, can lead to the inference that outward characteristics exhibited by members of any biologically relevant category—be it cells, species, or types of ecosystems—should be relatively uniform, static, and predictable. Consequences of essentialist thinking such as assumptions about homogeneity and stability have been widely discussed as impediments to both individual understanding of biological concepts (e.g., Shtulman and Schulz, 2008; Coley and Muratore, 2012; Gelman and Rhodes, 2012) and the progress of biological science as a discipline.

### Table 1. Misconception challenge statements used in the present study, along with hypothesized related cognitive construal

<table>
<thead>
<tr>
<th>Hypothesized related cognitive construal</th>
<th>Biological misconception statement</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teleological thinking</td>
<td>Plants produce oxygen so that animals can breathe.</td>
<td>Wanderee, 1986; Stavy et al., 1987; Tamir, 1989; Anderson et al., 1990; Leach et al., 1992; Songer and Mintzes, 1994; Kuech et al., 2003; Özay and Ostas, 2003; Köse, 2008; Bishop and Anderson 1990; Passmore and Stewart, 2002; Stern and Roseman, 2004</td>
</tr>
<tr>
<td>Essentialist thinking</td>
<td>Apart from differences due to age and sex, members of the same species are essentially identical; any variability is biologically unimportant.</td>
<td>Greene, 1990; Anderson et al., 2002; Passmore and Stewart, 2002; Shtulman, 2006; Gelman and Rhodes, 2012</td>
</tr>
<tr>
<td>Anthropocentric thinking</td>
<td>Homeostasis keeps the body static and unchanging.</td>
<td>Westbrook and Marek, 1992</td>
</tr>
<tr>
<td>Anthropocentric thinking</td>
<td>Apart from differences due to age and sex, members of the same species are essentially identical; any variability is biologically unimportant.</td>
<td>Hackling and Treagust, 1984</td>
</tr>
<tr>
<td>Anthropocentric thinking</td>
<td>Different cells in an organism (e.g., skin, muscle, nerve) contain different DNA.</td>
<td>D’Avanzo 2003</td>
</tr>
<tr>
<td>Anthropocentric thinking</td>
<td>Without outside influences, ecological communities will remain stable indefinitely.</td>
<td>D’Avanzo 2003</td>
</tr>
<tr>
<td>Anthropocentric thinking</td>
<td>Humans have caused the majority of extinctions.</td>
<td>AAAS, 2014b</td>
</tr>
<tr>
<td>Anthropocentric thinking</td>
<td>Plants get their food from the soil.</td>
<td>Stavy et al., 1987; Tamir, 1989; Anderson et al., 1990; Leach et al., 1992; Songer and Mintzes, 1994; Wanderee, 1986; Kuech et al., 2003; Özay and Ostas, 2003; Köse, 2008</td>
</tr>
<tr>
<td>Anthropocentric thinking</td>
<td>The heart decides how much blood is needed throughout the body and adjusts the rate at which it beats accordingly.</td>
<td>Inagaki and Hatano, 2002; Morris et al., 2000; Miller and Bartsch, 1997</td>
</tr>
<tr>
<td>Anthropocentric thinking</td>
<td>Competition between organisms involves direct, aggressive interaction.</td>
<td>AAAS, 2014a</td>
</tr>
</tbody>
</table>
(e.g., Hull, 1965; Mayr, 1982). We hypothesize that essentialist thinking may strongly influence our intuitive understanding of biological entities and systems, as well as species. The second section of Table 1 lists examples of misconceptions documented by multiple research groups ranging from genetics to ecology. We hypothesize that essentialist thinking may underlie all of these misconceptions, including assumptions that simple one-to-one correspondence exists between essence (DNA) and observable properties, assumptions about species-wide homogeneity, and the view that ecological systems are static. Thus, like teleological thinking, essentialist thinking may provide an underlying explanation for a variety of seemingly unrelated biological misconceptions.

Hypothesized Anthropocentric Misconceptions. Anthropocentric thinking is a natural manifestation of our powerful analogical reasoning abilities. However, when applied to the life sciences, this can lead students to a distorted understanding of both humans and other species. The third section of Table 1 lists examples of diverse misconceptions that we hypothesize may all share an underlying origin in anthropocentric thinking. In these examples, anthropocentric thinking can lead to distortions of humans’ role in the biological world, overattribution of human (or animal) functions to dissimilar organisms (e.g., plants), or personification of physiological processes. Similar to teleological and essentialist thinking, anthropocentric thinking may represent a third cognitive construal that underlies this set of seemingly unrelated biological misconceptions.

Testing Hypotheses about Relations between Cognitive Construals and Biological Misconceptions

As detailed above, we hypothesize that many common biological misconceptions may stem not from the complexity or opacity of the concepts themselves, but because they may arise from informal, intuitive, and deeply held ways of understanding the world (Coley and Tanner, 2012). Little is known about the presence and, more importantly, the relations between intuitive biological reasoning and discipline-based biological reasoning among young adults beginning their university studies. To address this gap in our collective knowledge, we investigated the following research questions among entering biology majors and entering students in majors outside the life sciences (henceforth, “nonmajors”).

What Is the Prevalence of Hypothesized Construal-Based Misconceptions among Undergraduate Students, and How Do Biology Majors and Nonmajors Differ? We hypothesize that biological novices (both majors and nonmajors) will show clear evidence of the specific biological misconceptions described earlier (see Table 2). Incoming biology majors may show a somewhat lower prevalence of misconceptions due to more biology and related course work in high school or more previous interest-driven exploration of the topic.

What Is the Prevalence of Construal-Based Reasoning among Undergraduate Students, and How Does This Differ between Biology Majors and Nonmajors? We hypothesize that, because cognitive construals are an integral component of intuitive biological reasoning, both groups of biological novices (entering biology majors and nonmajors alike) will show clear evidence of cognitive construals in their explanations of biological ideas. Incoming biology majors may show a somewhat lower prevalence of construals in this context, in particular if their experiences in formal secondary education resulted in decreased application of intuitive construals to discipline-based problems.

How Are Construal-Based Reasoning and Biological Misconceptions Related, and Are Those Relations the Same for Biology Majors and Nonmajors? To the extent that intuitive cognitive construals underlie biological misconceptions, we hypothesize that the presence of construals will be positively associated with the strength and prevalence of hypothesized construal-based misconceptions. Moreover, we expect associations to be exclusive, such that, for example, teleological construals should be associated with hypothesized teleological misconceptions but not necessarily with hypothesized essentialist or hypothesized anthropocentric misconceptions. Finally, we hypothesize that these relations will hold equally for both biological novices who are entering biology majors and who are entering nonmajors, unless prior formal biology education has somehow mitigated these connections.

We present here an integrated research study that brings together the theoretical frameworks, research methodologies, and analytical approaches of both cognitive psychology and science education to test these hypotheses.

METHOD

Participants

A total of 137 undergraduate students at Northeastern University participated in the study. One group, which we will refer to as “biology majors” (n = 69) were first-year students enrolled in two sections of a seminar designed for freshman biology, biochemistry, and behavioral neuroscience majors with Advanced Placement (AP) biology credit. These students had all taken AP biology courses in high school and achieved a score of 4 or 5 on the AP Biology exam. These students participated as part of their normal classroom activities. The “nonmajors” group (n = 68) was composed of students enrolled in an introductory psychology course who participated to partially fulfill a course research hours requirement. Nonmajors were prescreened to ensure that they met two criteria. The first was that they were not majoring in biology, biochemistry, or behavioral neuroscience (nonmajors represented a wide variety of majors, most of which were in the humanities and social sciences) and the second was that they had earned AP credit in some subject other than biology. This was to roughly equate general academic ability in the two groups.

Biological Misconception Statements

To assess the spontaneous use of teleological thinking, essentialist thinking, and anthropocentric thinking in reasoning about biological science, we constructed 12 misconception statements. Four of these were hypothesized to be associated with each of the three construals. (These misconception statements are listed in Table 1.) Misconception statements

Vol. 14, Highlights of 2015 129
were chosen to represent the five core concepts articulated in the Vision and Change in Undergraduate Biology initiative (AAAS, 2011): 1) evolution (the diversity of life forms that have evolved over time through mutations, selection, and genetic change); 2) structure and function (the basic units of biological structures that define the functions of all living things); 3) information flow, exchange, and storage (the influence of genetics on the control of the growth and behavior of organisms); 4) pathways and transformations of energy and matter (the ways in which chemical transformation pathways and the laws of thermodynamics govern the growth and change of biological systems); and 5) systems (the ways in which living things are interconnected and interact with one another).

We also chose statements that we hypothesized to be closely linked to one of the three cognitive construals discussed in Coley and Tanner (2012): teleological, essentialist, or anthropocentric thinking. Specifically, teleological statements were chosen to represent misconceptions based on the idea that an outcome, purpose, or goal plays a causal role in a biological process. Essentialist statements involved misconceptions based on several related ideas: that shared underlying essence results in homogeneity, that bodily systems and ecological communities are static unless perturbed, and that surface/functional differences should imply the presence of underlying differences. Finally, anthropocentric statements reflected misconceptions based on false analogies to animals, overestimates of the impact of humans across geological time, and personified views of organ function and competition.

We split the misconception statements into two comparable assessment forms of six statements each (two statements corresponding to each construal) and presented two statements, with different hypothesized underlying construals, per page. Each participant responded to one assessment form only and post hoc analysis suggested no systematic differences between forms. For each statement, participants indicated whether they agreed with the statement on a scale (1 = “strongly disagree,” 2 = “disagree,” 3 = “agree,” 4 = “strongly agree,” 5 = “don’t know”); this portion of their assessment responses will be referred to throughout as their “agreement with the biological misconception.” Participants were additionally asked to provide detailed written explanations of their responses; this portion of their assessment responses will be referred to throughout as their “written reasoning about the biological misconception.”

Data Collection Procedure

Biology majors were assessed during a full class period at the beginning of the academic year. The assessment form was distributed, and students were told to take their time and answer carefully. Nonmajors were assessed individually or in small groups in a laboratory setting and were given the same instructions. Misconception statements were presented in the context of a larger assessment tool that also included several measures of intuitive biological thinking drawn from the cognitive science literature. Misconception statements were always presented first to avoid any influence of the intuitive biology thinking measures on responses to the misconception statements. Completion of the entire assessment typically took 45–60 min. Only analyses of the misconception statements are presented here.

Scoring and Statistical Analyses

Assessing Presence of Misconceptions through Analysis of Agreement. To address our first research question, “What is the prevalence of biological misconceptions among undergraduate students, and how do biology majors and nonmajors differ?,” we examined students’ agreement or disagreement with the misconception statements as indicated by their responses to the five-point scale. We scored this in two ways.

Categorical Agreement. First, we examined agreement as a binary categorical variable. To do so, we classified each student as either “agreeing” (i.e., giving a rating of 3 or 4) or “disagreeing” (giving a rating of 1 or 2) with each individual misconception statement and compared the number of biology majors and nonmajors who agreed with zero, one, or two misconception statements of each type (teleological, essentialist, anthropocentric) via 2 (major) × 3 (0/1/2 agreements with misconception statements) chi-square analyses. To look at differences between majors and nonmajors in agreement with individual misconception statements, we conducted 2 (major) × 2 (agree/disagree) chi-square tests separately for each of the 12 statements.

Degree of Agreement along a Continuum. Second, we examined participants’ level of agreement with each type of misconception statement as a continuous measure. To do so, we averaged the two agreement ratings for each type of statement (teleological, essentialist, anthropocentric), yielding three scores for each student for each statement type. Scores could range from one to four, with higher scores indicating stronger agreement with the statements, which in turn may suggest stronger misconceptions. To examine differences among different types of misconception statements and between majors and nonmajors, we conducted a 2 (major) × 3 (type of statement) mixed analysis of variance (ANOVA). All pairwise comparisons reported as significant are based on Bonferroni-corrected t tests with \( p < 0.050 \).

Assessing Presence of Cognitive Construals through Analysis of Written Reasoning. To address our second research question, “What is the prevalence of construal-based reasoning in discipline-based biological reasoning, and how does this differ between biology majors and nonmajors?,” we coded students’ explanations for the presence of teleological, essentialist, and anthropocentric thinking. To do so, we transcribed each written explanation students provided for their position on each misconception statement. Six trained coders, blind to the participants’ major and the type of misconception statement, coded each response independently for the presence or absence of each construal (teleological, essentialist, and anthropocentric thinking). Thus, different construals could be identified within a single response. Coding criteria are summarized in Table 2, along with examples of responses for which coding decisions were unanimous. Because each challenge statement has important individual characteristics, the precise way that we applied the generic coding categories varied slightly from item to item; details are available from the authors upon request.

For the purpose of analyses, a response was considered to embody a construal when a majority of independent coders (i.e., four or more out of six) indicated the presence of
INTUITIVE THINKING AND MISCONCEPTIONS

**Table 2. Coding criteria for identifying cognitive construals in written explanations with examples from student responses**

<table>
<thead>
<tr>
<th>Cognitive construal</th>
<th>Coding criteria/definition</th>
<th>Examples of student language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teleological thinking</td>
<td>Response includes one or more explanations of biological structures, processes, or phenomena by reference to their supposed purpose, goal, function, or outcome.</td>
<td>“Adaptations are made in order to promote reproduction and the continuation of that particular species.” (biology major) “Plants produce oxygen for all kinds of life forms to help fuel the reactions necessary for daily life.” (nonmajor)</td>
</tr>
<tr>
<td>Essentialist thinking</td>
<td>Response includes one or more explanations of biological structures, processes, or phenomena consistent with the idea that underlying shared properties cause external features, and that the outward characteristics exhibited by members of any biologically relevant category—be it cells, species, or types of ecosystems—should be relatively uniform, static, and predictable. This includes indirect reference to a biological category or group that implies uniformity with respect to a property or behavior via generic language, e.g., “cats eat mice.”</td>
<td>“The coding for each type of cell is different so that each cell has a unique function.” (biology major) “Nature has a delicate balance, so without any drastic changes, ecological communities will mostly remain stable.” (nonmajor)</td>
</tr>
<tr>
<td>Anthropocentric thinking</td>
<td>Response explains biological structures, processes, or phenomena by comparison with or analogy to humans or by mentioning humans, their roles, or their interventions. Response includes the inappropriate assignment of human (or animate) characteristics to nonhuman (or inanimate) entities.</td>
<td>“The heart does not regulate itself but is instructed by the brain. The brain recognizes either an increase or decrease in the levels of oxygen + carbon dioxide in the body + adjusts the heart rate accordingly.” (biology major) “Much like the nutrients humans and animals receive, plants get their nutrients from the soil. While not everything they need is in the soil, many are.” (nonmajor)</td>
</tr>
</tbody>
</table>

that construal in that response. To assess the prevalence of each cognitive construal in students’ biological reasoning, we compared the number of biology majors versus nonmajors who used each construal at least once across all six of their written explanations. We also conducted a 2 (major) × 3 (construal) × 2 (form) mixed ANOVA on the number of explanations (out of six) in which each construal was used by each student. We included “form” as a variable, because the two different forms of the assessment (each administered to roughly half of the participants) contained different misconception statements. We followed up this ANOVA with independent-samples *t* tests and univariate ANOVAs run separately for biology majors and nonmajors, for which all pairwise comparisons reported as significant are based on Bonferroni-corrected *t* tests with *p* < 0.050.

**Assessing Relations between Misconceptions and Cognitive Construals.** We tested our third research question, “How are construal-based reasoning and misconceptions related, and are those relations the same for biology majors and nonmajors?,” in two ways. First, in order to examine relations between individual misconceptions and the construals we hypothesized to be related to them, for each misconception statement, we classified each participant based on 1) whether they agree or disagreed with the statement (as an indication of whether they held the misconception) and 2) whether the construal we hypothesized to be associated with that misconception was present in their written explanation or not. We then conducted 2 × 2 chi-square analyses on these data separately for biology majors and nonmajors and combined for all students. Second, in order to examine more general relations between frequency of types of misconceptions and the construals we hypothesized to be associated with them, we ran multiple regression analyses. We used the frequency of each cognitive construal (teleological, essentialist, and anthropocentric thinking) across all written explanations as predictor variables and ran three different regressions, each of which used the average agreement with each specific type of misconception statement as an outcome variable. We performed these analyses for all students and also separately for biology majors and nonmajors.

**RESULTS**

We organize the results to address the three research questions raised previously. First, what is the prevalence of biological misconceptions among undergraduate students, and how do biology majors and nonmajors differ? Second, what is the prevalence of intuitive cognitive construals in discipline-based biological reasoning, and how does this differ between biology majors and nonmajors? Third, how are construal-based reasoning and misconceptions related, and are those relations the same for biology majors and nonmajors?

**Prevalence of Biological Misconceptions Among Biology Majors and Nonmajors**

**Categorical Agreement.** All together, 93% of biology majors agreed with at least one misconception statement, as did 98% of nonmajors. Individual misconception statements varied widely in the degree to which students agreed with them, ranging from a high of 87% of students agreeing that “Species adapt to their environment in order to survive” to a low of 14% of students agreeing that “Different cells in an organism (e.g., skin, muscle, nerve) contain different DNA.” Recall that each student responded to two misconception statements of each type (teleological, essentialist, anthropocentric). The percentage of biology majors and
J. D. Coley and K. Tanner

nonmajors who agreed with zero, one, or two statements of each type is depicted in Figure 1. Students rarely selected the “don’t know” option on the agreement scale (M = 0.28 times out of six opportunities). Not surprisingly, such responses were more frequent for nonmajors (M = 0.43) than for biology majors (M = 0.14, t(135) = 2.79, p = 0.006). These responses were excluded from further analyses and represent < 5% of all responses. In the sections following, we examine agreement separately for each type of misconception statement.

Teleological Misconceptions. As depicted in Figure 1, the number of students agreeing with zero, one, or two statements did not vary by major for teleological misconceptions ($\chi^2(2,137) = 0.68, p = 0.711$). As seen in Figure 2, three of the four teleological misconception statements were endorsed by more than 50% of students. Indeed, more than 75% of students from both groups agreed with teleological statements having to do with evolution (“Species adapt to their environment in order to survive” and “Many species develop protective camouflage to avoid predators”). Majors and nonmajors did not differ on these statements but did so on the remaining two. Specifically, more nonmajors than majors agreed that “Plants produce oxygen so that animals can breathe” ($\chi^2(1,67) = 8.80, p = 0.003$), whereas more biology majors than nonmajors agreed that “Genes turn on so that a cell can develop properly” ($\chi^2(1,58) = 6.28, p = 0.012$).

Essentialist Misconceptions. In Figure 1, we see evidence that nonmajors were more likely to agree with essentialist misconception statements than biology majors. Specifically, nonmajors were more likely than majors to agree with one misconception statement (43 vs. 23%), whereas biology majors were more likely than nonmajors to agree with none of the misconception statements (67 vs. 39%, $\chi^2(2,136) = 10.60, p = 0.005$). Agreement with individual statements is depicted in Figure 3, which shows that the only essentialist misconception statement garnering more than 50% agreement was “Without outside influences, ecological communities will remain stable indefinitely.”

Figure 1. Percentage of students agreeing with zero, one, or two teleological, essentialist, or anthropocentric misconception statements.

Figure 2. Percentage of students agreeing with each teleological misconception statement. Note: biology majors and nonmajors differ via chi-square test: *, $p < 0.05$; **, $p < 0.01$. 

132 CBE—Life Sciences Education
Intuitive Thinking and Misconceptions

Agreement with other essentialist statements was relatively low; however, consistent differences between biology majors and nonmajors emerged. More nonmajors than majors agreed with the “ecological communities” statement ($\chi^2(1,64) = 6.14$, $p = 0.013$), and more nonmajors also agreed that “members of the same species are essentially identical; any variability is biologically unimportant” ($\chi^2(1,68) = 6.37$, $p = 0.012$). Marginally more nonmajors than majors agreed that “homeostasis keeps the body static and unchanging” ($\chi^2(1,57) = 3.19$, $p = 0.074$).

**Anthropocentric Misconceptions.** Nonmajors were also more likely than biology majors to agree with anthropocentric misconception statements (see Figure 1). Specifically, nonmajors were more likely than majors to agree with both statements (31 vs. 9%), whereas biology majors were again more likely than nonmajors to agree with none of the misconception statements (51 vs. 28%, $\chi^2(2,136) = 13.07$, $p = 0.001$). Response patterns for individual statements are depicted in Figure 4, which shows that a majority of students (including 49% of biology majors) agreed that “Plants get their food from the soil.” Although nonmajors agreed more than biology majors with each anthropocentric statement, this difference was only statistically reliable for “Competition between organisms involves direct, aggressive interaction” ($\chi^2(1,67) = 6.67$, $p = 0.010$).

**Degree of Agreement along a Continuum.** As depicted in Figure 5, ANOVA showed that average agreement ratings were highest for teleological statements, intermediate for anthropocentric statements, and lowest for essentialist statements, $F(2,268) = 69.26$, $p < 0.001$, post hoc pairwise $p < 0.05$. Moreover, mean agreement was higher among nonmajors ($M = 2.60$) than among biology majors ($M = 2.21$, $F(1,134) = 23.63$, $p < 0.001$); this was true for all three types of misconception statements. The interaction was not significant.

**Prevalence of Cognitive Construals in Written Explanations of Biology Majors and Nonmajors**

Cognitive construals were used by a majority of students in their written explanations for their positions on the misconception statements; the percentage of students who used each construal at least once in their written responses to the six misconception statements is depicted in Figure 6. Altogether, 58% of students used teleological reasoning at least once, 88% used essentialist reasoning at least once, and...
68% used anthropocentric reasoning at least once. Biology majors and nonmajors did not differ in the frequency with which they used any construal at least once in their written explanations ($\chi^2(1,136) < 1.00, p > 0.420$).

When we analyzed the mean number of times each student used each type of construal, ANOVA revealed that essentialist construals were more common than anthropocentric construals, which in turn were more common than teleological construals ($F(2,270) = 3.63, p = 0.028$). However, this pattern differed for biology majors versus nonmajors, as indicated by a significant interaction between construal and major ($F(2,264) = 3.72, p = 0.025$). As can be seen in Figure 7, for biology majors, essentialist construals were more common than teleological or anthropocentric construals. In contrast, for nonmajors, essentialist and anthropocentric construals were more common than teleological construals. Anthropocentric construals were more common for nonmajors than for biology majors ($t(135) = 2.12, p = 0.036$), but differences for other construals were not statistically reliable.

There were no differences in evidence for cognitive construal-based reasoning in the written responses to the two forms of the assessment.

### Relations Between Biological Misconceptions and Cognitive Construals

#### Relations between Individual Misconceptions and Construals.

If agreement with a particular misconception stems from a specific underlying cognitive construal, then students who hold that misconception should be more likely to use that construal in their written explanations than students who do not hold that misconception. For eight of the 12 misconception statements we examined, this prediction was supported. Results of the chi-square analyses for these individual statements are presented in Table 3.

#### Hypothesized Teleological Misconceptions.

When responses to individual misconception statements were considered, agreement was associated with teleological reasoning for three of the four hypothesized teleological misconceptions (see Table 3). As can be seen in Figure 8, for these three items, students who agreed with the misconception statement were more likely to explain their position in a way that demonstrated teleological reasoning than students who disagreed with the statement. Results for the fourth item—“genes turn on”—although not statistically significant, still trended in the same direction.

#### Hypothesized Essentialist Misconceptions.

Agreement with all four individual hypothesized essentialist misconception statements was associated with essentialist reasoning (see Table 3). As can be seen in Figure 9, students who agreed with the misconception statement were more likely to explain their position in a way that demonstrated essentialist reasoning than students who disagreed with the statement.

#### Hypothesized Anthropocentric Misconceptions.

Agreement and anthropocentric reasoning were only associated for one hypothesized anthropocentric item, “humans cause extinctions” (see Table 3). Nevertheless, as can be seen in Figure 10, for this item, as well as the “plants get food” item ($p = 0.149$), students who agreed with the misconception statement were more likely to explain their position in a way that demonstrated anthropocentric reasoning than students who disagreed with the statement.

### Relations between Types of Misconceptions and Construals.

If the extent to which students hold a type of misconception
Table 3. Results of chi-square analyses for item-wise agreement-by-construal comparisons for biology majors, nonmajors, and all students combined.

<table>
<thead>
<tr>
<th>Hypothesized underlying construal</th>
<th>Misconception</th>
<th>All students</th>
<th>Biology majors</th>
<th>Nonmajors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teleological thinking</td>
<td>Camouflage</td>
<td>***</td>
<td>+</td>
<td>**</td>
</tr>
<tr>
<td></td>
<td>Species adapt</td>
<td>*</td>
<td>*</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Plants produce O₂</td>
<td>*</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Genes turn on</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Essentialist thinking</td>
<td>Homeostasis</td>
<td>*</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Communities remain stable</td>
<td>***</td>
<td>*</td>
<td>**</td>
</tr>
<tr>
<td></td>
<td>Cells have different DNA</td>
<td>***</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td></td>
<td>Members are identical</td>
<td>***</td>
<td>+</td>
<td>***</td>
</tr>
<tr>
<td>Anthropocentric thinking</td>
<td>Humans cause extinction</td>
<td>**</td>
<td>*</td>
<td>**</td>
</tr>
<tr>
<td></td>
<td>Plants get food</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Heart decides</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Competition</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

*p > 0.10.  
*p < 0.10.  
*p < 0.05.  
*p < 0.01.  
*p < 0.001.

Figure 8. Percentage of students who showed evidence of teleological construals in their written explanations as a function of whether they agreed or disagreed with the misconception statement.
Table 4). In contrast, regression analyses revealed no relation between agreement with hypothesized anthropocentric misconception statements and frequency of any type of cognitive construal, for biology majors or nonmajors (see Table 4).

**DISCUSSION**

Research and theory development in cognitive psychology and science education research remain largely isolated. Biology education researchers have documented persistent scientifically inaccurate ideas, often termed *misconceptions*, among biology students in multiple biological domains. In parallel, cognitive and developmental psychologists have described intuitive conceptual systems—teleological, essentialist, and anthropocentric thinking—that humans use to reason about biology. We have hypothesized that seemingly unrelated biological misconceptions may have common origins in these intuitive ways of knowing, termed *cognitive construals* (Coley and Tanner, 2012). In this paper, we investigated the following questions: 1) What is the prevalence...
of construal-based misconceptions among undergraduate students, and how do biology majors and nonmajors differ? 2) What is the prevalence of construal-based reasoning in discipline-based biological reasoning, and how does this differ between biology majors and nonmajors? 3) How are construal-based reasoning and misconceptions related, and are those relations the same for biology majors and nonmajors? In the following sections, we explore the implications of our results for each of these research questions in the contexts of formal university biology education, biology education research, and cognitive and developmental psychology, as well as potential future research directions.

What Is the Prevalence of Biological Misconceptions among Undergraduate Students, and How Do Biology Majors and Nonmajors Differ?

While many assessment tools have been developed at the college level to diagnose biological misconceptions, there has been less systematic investigation of the nature and origins of those misconceptions that persist among science majors in higher education. To assess the presence of biological misconceptions, we simply asked participants whether they agreed or disagreed with a series of biological misconception statements. Our results suggest that the vast majority of undergraduates—biology majors (93%) and nonmajors (98%) alike—agreed with at least one biological misconception. Overall, nonmajors were more likely to agree with misconception statements than biology majors, and this difference was most pronounced for misconceptions hypothesized to correspond to essentialist and anthropocentric thinking. However, what is perhaps most striking is the similarity between majors and nonmajors. For six of 12 misconception statements, the two groups did not differ statistically on the likelihood of agreeing with the statement. And although agreement ratings were higher on average for nonmajors than for majors, both groups showed the same relative ordering—agreeing most with teleological misconceptions, followed by anthropocentric and finally essentialist misconceptions.
Table 4. Relations between agreement with biological misconception statements and presence of cognitive construals in written explanations

<table>
<thead>
<tr>
<th>Agreement with misconception statements</th>
<th>Frequency of cognitive construals in written explanations*</th>
<th>$r^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Teleological thinking</td>
<td>Essentialist thinking</td>
</tr>
<tr>
<td>All students: Agreement with...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teleological misconceptions</td>
<td>0.224**</td>
<td>0.086</td>
</tr>
<tr>
<td>Essentialist misconceptions</td>
<td>−0.027</td>
<td>0.283***</td>
</tr>
<tr>
<td>Anthropocentric misconceptions</td>
<td>0.057</td>
<td>0.190</td>
</tr>
<tr>
<td>Biology majors: Agreement with...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teleological misconceptions</td>
<td>0.309*</td>
<td>0.123</td>
</tr>
<tr>
<td>Essentialist misconceptions</td>
<td>0.190</td>
<td>0.319**</td>
</tr>
<tr>
<td>Anthropocentric misconceptions</td>
<td>0.119</td>
<td>0.195</td>
</tr>
<tr>
<td>Nonmajors: Agreement with...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teleological misconceptions</td>
<td>0.141</td>
<td>0.032</td>
</tr>
<tr>
<td>Essentialist misconceptions</td>
<td>−0.174</td>
<td>0.176</td>
</tr>
<tr>
<td>Anthropocentric misconceptions</td>
<td>0.091</td>
<td>0.061</td>
</tr>
</tbody>
</table>

For ease of interpretation, any entries in the table for which $p \leq 0.05$ appear in bold.

* $p \leq 0.05$.
** $p \leq 0.01$.
*** $p \leq 0.001$.

Entries represent standardized regression coefficients ($\beta$).

Because the biological misconceptions used here represent only a small subset of possible misconceptions in these subject areas, differences among types of statements should be interpreted with caution. Although we chose the misconception statements to correspond to different underlying cognitive construals, we made only informal attempts to equalize the a priori plausibility of the statements, and we did not draw the statements randomly from a larger sample of potential misconceptions. Therefore, we cannot claim that, for example, misconceptions based on teleological thinking are stronger in general than misconceptions based on essentialist thinking; this may be an artifact of the misconceptions we examined.

Although the proliferation of concept inventories and the current dialogue about the nature of misconceptions suggest that instructors in higher life sciences education are attending to misconceptions, these results demonstrate the importance of further research on the nature and origins of misconceptions and their roots in intuitive student reasoning. The small but consistent difference in misconceptions between incoming biology majors and nonmajors suggests that advanced high school course work in biology and/or an underlying interest in and facility with the life sciences have some effect on biological misconceptions. Nevertheless, our data document persistent biological misconceptions among both majors and nonmajors. This contradicts a tacit assumption that emerging experts, namely university-level biology majors, do not hold the same misconceptions that have been previously documented among younger students, K–12 teachers, nonmajors, and the general public.

**What Is the Prevalence of Construal-Based Reasoning among Undergraduate Students, and How Does This Differ between Biology Majors and Nonmajors?**

While extensive research on construal-based reasoning has been conducted in young children, investigation of the “end-state” of conceptual development in older and presumably more advanced individuals such as undergraduate students has received much less attention (Coley, 2000). We identified instances of construal-based reasoning in participants’ explanations for their agreement or disagreement with the misconception statements based on consensus among multiple, independent, trained coders. Results show that a majority of students explicitly used each type of construal (teleological, essentialist, and anthropocentric thinking) in their written explanations on at least one occasion. Although the mean frequencies for each construal may appear low (see Figure 7), our coding system likely underestimates the use of intuitive construal-based reasoning among college students, for at least two reasons. First, such construals are typically implicit, and our coding system captured only explicit construal-based reasoning (see Taber and Garcia-Franco, 2010, for a discussion of the role of implicit knowledge in students’ scientific explanation). Second, we utilized a relatively conservative criterion (4/6 independent coders) for identifying instances of each construal. Thus, we conclude that intuitive construal-based reasoning is readily observable among entering university students.

Indeed, we observed few differences in construal-based reasoning between biology majors and nonmajors, documenting that intuitive construal-based reasoning is readily observable in the written responses of both of these populations. This fits with other recent work showing evidence of intuitive biases in adult reasoning (e.g., Shutlman and Valcarcel, 2012; Kelemen et al., 2013; Eidson and Coley, 2014). It also fits with previous demonstrations of students’ spontaneous and explicit use of intuitive cognitive construals in discipline-based biology education research literature (e.g., Tamir and Zohar, 1991; Friedler, et al., 1993; Zohar and Ginossar, 1998; Nehm and Ridgeway, 2011). For the fields of cognitive and developmental psychology, which have focused largely on such reasoning among elementary school
children, these findings clearly demonstrate the persistence of construal-based intuitive biological reasoning into young adulthood.

Even more striking was the lack of differences between biology majors and nonmajors in their use of construal-based reasoning in discipline-based biology problems. Indeed, differences in misconceptions, albeit small, were systematic, whereas differences in construal-based reasoning were virtually nonexistent in these investigations. Although the biology majors were all first-semester undergraduates, and therefore at the very beginning of a university-level life sciences curriculum, they had successfully navigated a high school–level science curriculum rigorous enough to enable them to score well on the standardized AP test. These results may seem counterintuitive, because one might hypothesize that greater experience in formal education in biology might have caused biology majors to abandon intuitive ways of reasoning, especially about explicitly biological content. Our results suggest otherwise.

How Are Construal-Based Reasoning and Misconceptions Related, and Are Those Relations the Same for Biology Majors and Nonmajors?

Our most critical hypotheses were that the presence of construal-based reasoning would be positively associated with the strength and prevalence of misconceptions and that such associations would be exclusive and construal specific. For example, we hypothesized that the presence of teleological construal-based reasoning would be associated with agreement with hypothesized teleological misconceptions but not necessarily with hypothesized essentialist or hypothesized anthropocentric misconceptions. Our results support these predictions. For three of four teleological misconception statements, agreement with the misconception was associated with students’ use of explicit teleological reasoning in their written explanations. Likewise, for all four of the essentialist misconception statements, agreement with the misconception was associated with students’ use of explicit essentialist reasoning in their written explanations. These results are consistent with previous work showing that misconceptions about evolution are associated with essentialist thinking (e.g., Shtulman and Schulz, 2008) and teleological thinking (e.g., Klemen and Rosset, 2009). They also extend these findings by demonstrating a linkage between systems of misconceptions that transcend traditional biological subject areas and underlying intuitive cognitive construals. This pattern of results in turn supports our primary hypothesis.

When results were aggregated across responses to all items, we observed a striking difference in the relations between construal-based reasoning and misconception agreement for biology majors compared with nonmajors. Specifically, for biology majors, we saw very specific and precise relations; the frequency of teleological construals in written explanations was associated exclusively with agreement with teleological misconception statements, and frequency of essentialist construals in written explanations was associated exclusively with agreement with essentialist misconception statements. In contrast, for nonmajors, the overall frequency of construals in written explanations was unrelated to their agreement with corresponding misconception statements. This raises the intriguing possibility that formal education in the biological sciences might actually serve to reify intuitive biological thought among biology majors, resulting in stronger and more specific relations between construal-based reasoning and agreement with related misconceptions than for nonmajors. As a reminder to the reader, biology majors were those students who had scored a 4 or 5 on the AP Biology exam in high school, as compared with nonmajors who had not, but who had achieved similarly on a non-biology AP exam. Unfortunately, no additional information is available about the high school biology experiences of students in this study. In future studies, collection of more detailed information on subjects’ high school biology instruction could inform the interpretation of results.

Interestingly, for hypothesized anthropocentric misconception statements, the relation between construal-based reasoning and misconception agreement was weaker. Agreement was associated with used of explicit anthropocentric reasoning on only one out of four misconception statements, and regression analysis revealed no relation between agreement and use of anthropocentric reasoning across all explanations. However, this does not appear to be due to a lack of agreement with the misconceptions hypothesized to be based in anthropocentric thinking nor due to a lack of instances of anthropocentric reasoning in students’ written statements. Rather, we simply observed much weaker interrelations between construal-based reasoning and misconception agreement for anthropocentric thinking than we did for teleological thinking or essentialist thinking. There are several possible explanations for this finding. Interactions between anthropocentric thinking and biological misconceptions might be fundamentally different from those for essentialist thinking and teleological thinking. Alternatively, if the origin of these relations is in formal biology education, perhaps anthropocentric reasoning is less often used in making complex biological ideas accessible to novices and therefore less strongly linked to misconceptions. Perhaps our anthropocentric challenge statements were nonrepresentative; Talanquer (2010, 2013), for instance, finds teleological thinking on a very specific subset of chemistry problems (i.e., those having to do with osmotic flow but not with freezing and boiling points of solutions). Or perhaps our coding system was insufficiently sensitive to anthropocentric thinking. Further research will be needed to sort out these possibilities.

Future Research Directions: Origins, Persistence, and Generality

The findings presented above represent an initial foray into exploring the interactions between intuitive ways of knowing from psychology and misconceptions observed in formal biology education. In line with our hypotheses, we have demonstrated specific linkages between construal-based intuitive reasoning and particular sets of biological misconceptions. We have also shown that relations between intuitive reasoning and misconceptions differ for biology majors versus nonmajors. These findings raise questions about the origins, persistence, and generality of relations between intuitive reasoning and biological misconceptions. We explore below three lines of potential future research directions.

With respect to the origins, our evidence suggests that relations between construal-based reasoning and biological
misconceptions are stronger and more specific among biology majors than nonmajors. This difference raises the question of why those students with presumably more disciplinary interest in biology and perhaps more formal education in biology would be more likely to employ specific construal-based reasoning (e.g., essentialist reasoning) in their biological explanations of specific misconceptions (e.g., hypothesized essentialist misconceptions). One hypothesis is that formal secondary biology education itself may be somehow either driving or reifying these relations. If this were the case, one might hypothesize that construal-based reasoning would be found in the language of instruction among high school biology teachers. While it would be unexpected for high school biology teachers to exhibit the same relations as the entering biology majors studied here, these same individuals may unknowingly employ teleological and essentialist reasoning in their attempts to make complex biological ideas accessible to high school students. Multiple lines of previous research have investigated the extent to which teachers themselves may embrace common biological misconceptions (e.g., Nehm and Schonfeld, 2007). Additionally, Sadler and colleagues (2013) have documented that those teachers who are best able to predict students’ misconceptions and inaccurate reasoning are those teachers who are able to promote the largest learning gains for students. An investigation of how high school biology teachers perform on the assessments used here could clarify whether they also exhibit specific relations between construal-based reasoning and misconception agreement. Additionally, analysis of transcripts of the language used to teach those ideas related to the most agreed-with teleological (e.g., “Many species develop protective camouflage to avoid predators”) and essentialist misconceptions (e.g., “Apart from differences due to age and sex, members of the same species are essentially identical”) might reveal the use of construal-based language in biology teaching by high school instructors, even if they themselves do not endorse these relationships when assessed.

With respect to persistence, one wonders how formal undergraduate biology education will affect these specific relations between misconception agreement and construal-based reasoning. The findings presented here were only for those students who were beginning their university studies. If formal university biology education somehow disavows biology majors of biological misconceptions and the use of construal-based reasoning in biology, then we would hypothesize that, in a cross-sectional study, advanced or graduating biology majors would neither agree with the biological misconception statements used here nor employ construal-based reasoning in their explanations. Alternatively, if formal biology education is either not affecting or even driving these relations, we might hypothesize that biological misconception statements would persist among advanced or graduating biology majors and that the relations between their misconception agreement and construal-based reasoning might even increase in specificity (e.g., use of teleological reasoning in explaining hypothesized teleological misconception statements). Further, the persistence of these relations among advanced biology majors would make inquiry into the language used in college biology classrooms ripe for study. Similar to the high school biology teaching investigations proposed earlier, it would be unexpected for university biology instructors to exhibit the same relations as entering biology majors; however, they may unknowingly employ construal-based reasoning in the language of their teaching.

Finally, further research is needed to explore how extensive the relations between misconception agreement and construal-based reasoning really are with respect to a large set of misconception statements related to the three cognitive construals under study, and perhaps other cognitive construals as well. While these patterns were for the most part consistent for the misconception statements we examined, there may be other biological misconception statements that are rooted in multiple cognitive construals and that may therefore be simultaneously associated with several types of construal-based reasoning. To assess these and other possibilities, investigation of a larger set of randomly chosen misconceptions is important to extend the present findings and examine how robust these specific relations between misconceptions and construal-based reasoning are for biology majors.

CONCLUSIONS

In conclusion, the findings presented here represent a fruitful initial investigation of the interactions between intuitive ways of knowing long studied in cognitive psychology and misconceptions previously documented through biology education research. We found that the vast majority of university undergraduates in this study exhibited agreement with one or more biological misconception statements, with nonmajors being slightly more likely to agree than biology majors. In addition, both biology majors and nonmajors commonly used explicit teleological, essentialist, and anthropocentric reasoning in their explanations of misconception statements, with few differences observed between the two populations. Finally, and most importantly, our findings show very specific patterns of association between construal-based reasoning and biological misconceptions; moreover, these relations were stronger and more specific among biology majors than nonmajors. Taken together, these results complement previous findings in cognitive science and biology education research. They also extend these findings by demonstrating a linkage between systems of misconceptions that transcend traditional biological subject areas and underlying intuitive cognitive construals. They support the hypothesis that at least some common biological misconceptions may stem not from the complexity or opacity of the concepts themselves but from informal, intuitive ways of understanding the world (Coley and Tanner, 2012). They also raise the alarming possibility that formal education in the biological sciences might actually serve to reify intuitive biological thought, resulting in stronger and more specific relations between construal-based reasoning and agreement with related misconceptions. Finally, these findings suggest that further systematic investigation of issues at the interface of cognitive psychology and biology education, more generally, are ripe for investigation.

ACKNOWLEDGMENTS

This research was supported by National Science Foundation CAREER Award #0954127 to K.D.T. and by a Northeastern University College of Science Seed Grant to J.D.C. We thank Melanie Arnson,
REFERENCES


American Association for the Advancement of Science (AAAS) (2011). Vision and Change in Undergraduate Biology Education: A Call to Action, Washington, DC.


Vol. 14, Highlights of 2015 141

Intuitive Thinking and Misconceptions


A High-Enrollment Course-Based Undergraduate Research Experience Improves Student Conceptions of Scientific Thinking and Ability to Interpret Data

Sara E. Brownell,* † Daria S. Hekmat-Scafe,* Veena Singla,* Patricia Chandler Seawell,* Jamie F. Conklin Imam,* Sarah L. Eddy,‡ Tim Stearns,*§ and Martha S. Cyert*§

*Department of Biology, Stanford University, Stanford, CA 94305-5020; †Department of Biology, University of Washington, Seattle, WA 98195

Submitted May 30, 2014; Revised February 1, 2015; Accepted February 5, 2015
Monitoring Editor: James Hewlett

We present an innovative course-based undergraduate research experience curriculum focused on the characterization of single point mutations in p53, a tumor suppressor gene that is mutated in more than 50% of human cancers. This course is required of all introductory biology students, so all biology majors engage in a research project as part of their training. Using a set of open-ended written prompts, we found that the course shifts student conceptions of what it means to think like a scientist from novice to more expert-like. Students at the end of the course identified experimental repetition, data analysis, and collaboration as important elements of thinking like a scientist. Course exams revealed that students showed gains in their ability to analyze and interpret data. These data indicate that this course-embedded research experience has a positive impact on the development of students’ conceptions and practice of scientific thinking.

INTRODUCTION

Learning science means learning to do science.
—Vision and Change: A Call to Action (American Association for the Advancement of Science, 2011, p. 14)

Incorporating research experiences into the undergraduate curriculum is a major goal of national reform efforts (American Association for the Advancement of Science

© 2015 S. E. Brownell et al. CBE—Life Sciences Education © 2015 The American Society for Cell Biology. This article is distributed by the American Society for Cell Biology under license from the author(s). It is available to the public under an Attribution–Noncommercial–Share Alike 3.0 Unported Creative Commons License (http://creativecommons.org/licenses/by-nc-sa/3.0).

“ASCB®” and “The American Society for Cell Biology ®” are registered trademarks of The American Society for Cell Biology.

[AAAS], 2011; National Research Council [NRC], 2003). Research involving undergraduate students has been called the “purest form of teaching” (NRC, 2003, p. 87), and research experiences for undergraduates have been identified as “an integral component of biology education for all students” (AAAS, 2011, p. xiv). Despite these recommendations, the practical implications of involving students in research remain daunting. Historically, the integration of research into the undergraduate biology curriculum has primarily been in the form of research apprenticeships in faculty research labs (Russell et al., 2007), but there are not enough of these positions available at most institutions to give all students the opportunity to participate in authentic research.

One solution to this problem is to integrate research experiences into traditional high-enrollment lab courses (Handelsman et al., 2004; Sundberg et al., 2005; Auchincloss et al., 2014). Such course-based undergraduate research experiences, or CUREs, have five defining characteristics: 1) There is an element of discovery, so that students are working with novel data. 2) Iteration is built into the lab. 3) Students engage in a high level of collaboration. 4) Students learn scientific practices. 5) The topic is broadly relevant so that it could potentially be publishable and/or of interest to a group outside the class (Auchincloss et al., 2014). One of the major goals of these CUREs is that they reflect a real research experience in order to give students
a more accurate conception of how scientific research is done.

In response to calls for reform, a diverse range of CUREs have been developed (for a list, see www.curenet.franklin.uga.edu). Some of these courses are extensions of faculty members’ research projects (e.g., Brownell et al., 2012). In this format, students have the opportunity to work on a research problem in a formal course, and questions they explore can contribute to a faculty member’s research program (Kloser et al., 2011; Brownell and Kloser, 2015). An alternative model is a stand-alone lab course in which students conduct research that is only peripherally related, if at all, to a local faculty member’s research program. This approach is exemplified by the Howard Hughes Medical Institute (HHMI) SEA PHAGES program (Jordan et al., 2014) and the Genetics Education Partnership (Shaffer et al., 2010, 2014), both of which are multi-institution programs. These courses have been packaged so they can be implemented at a diverse range of institutions, and inter-institutional collaboration is encouraged.

While many CUREs currently exist, most of these have been small-sized classes taught to students who volunteer to participate. However, volunteer students and nonvolunteer students have previously been shown to have different affective gains from a CURE (Brownell et al., 2013), indicating that findings from volunteer populations may not be generalizable to students in required CUREs. Additionally, assessment of CUREs has been primarily in the form of student self-report surveys (e.g., CURE survey; Lopatto et al., 2008). While student self-reporting can be useful if one is interested in affective measures such as confidence or interest, it is not as effective at determining students’ abilities to interpret data or how similar their thinking processes are to expert scientists. Different means of assessment need to be used to further probe the impact of CUREs on students (Brownell and Kloser, 2015; Corwin et al., 2015).

We developed a CURE that replaced our “cookbook” introductory biology course; it is required of all undergraduate biology majors and is not directly related to any faculty member’s research at this institution. The primary purpose of this course is to engage students in research to shift their “thinking like a scientist” from novice to expert. This construct has been defined in previous literature (Druger et al., 2004; Hunter et al., 2007; Hurtado et al., 2009; Etkina and Planinši, 2014). We solicited responses from expert scientists who hold a PhD in their discipline to further articulate this construct; our definition is outlined in Table 1.

The purpose of this study was to assess the impact of this required high-enrollment CURE on 1) student conceptions of what it means to think like a scientist and 2) student ability to analyze and interpret scientific data, one of the key components of what it means to think like a scientist.

The novelty of this particular study is in the curriculum, how the course is required for a large population of introductory students, specific aspects of the course that promote and benefit from large-scale collaboration, and how we assessed the impact of the course curriculum through coded open-ended written responses and student learning gains in data interpretation. We hope that this can be a model for others interested in developing and assessing CUREs that are designed to give all students graduating with a biology degree the experience of doing research.

**CURRICULUM DESCRIPTION**

**Course Content: Investigating Human p53 Mutants Using Yeast as a Model System**

We sought to identify an unsolved scientific problem that would engage student interest in human biology and could be investigated using molecular and cell biology techniques accessible to students with no previous lab experience. We reasoned that analysis of a human disease-related protein would satisfy the first criterion and that use of budding yeast as an experimental system would satisfy the second. We chose the human tumor suppressor gene p53 as the basis for study. p53 is a transcription factor that promotes DNA repair, cell cycle arrest, and apoptosis (Levine, 1997; Sionov and Haupt, 1999). p53 is mutated in more than 50% of cancers (Hollstein et al., 1991; Soussi et al., 1994; Whibley et al., 2009). We note that Gammie and Erdeniz (2004) used a similar rationale in creating a smaller lab course based on the human mismatch repair protein MSH2 (Gammie and Erdeniz, 2004).

The specific scientific research question in our course is to characterize mutant versions of p53 identified in human tumors. Each mutant version of p53 contains a single point mutation in the p53 gene. Students are presented with a list of 10 human p53 mutations and are asked to design an experiment to determine the function of each mutant version (Brownell et al., 2009).

The novelty of this particular study is in the curriculum, how the course is required for a large population of introductory students, specific aspects of the course that promote and benefit from large-scale collaboration, and how we assessed the impact of the course curriculum through coded open-ended written responses and student learning gains in data interpretation. We hope that this can be a model for others interested in developing and assessing CUREs that are designed to give all students graduating with a biology degree the experience of doing research.

**Table 1.** Our definition of the construct “thinking like a scientist” based on prior studies and consensus of an expert panel of PhD-level scientists

<table>
<thead>
<tr>
<th>Element</th>
<th>Description</th>
<th>Agreement in prior literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Make discoveries</td>
<td>Scientists formulate questions, make observations, collect data, analyze and interpret data, test hypotheses, and draw conclusions.</td>
<td>Druger et al., 2004; Hunter et al., 2007; Hurtado et al., 2009; Etkina and Planinši, 2014</td>
</tr>
<tr>
<td>Make connections between seemingly unconnected phenomena</td>
<td>Scientists are able to think in multiple ways and design multiple types of experiments to test the same idea. New ideas often result from thinking differently. Science is not a linear process.</td>
<td>Hunter et al., 2007</td>
</tr>
<tr>
<td>Critically evaluate data with skepticism</td>
<td>Scientists critique both their own experiments and the experiments of others. There is the need to repeat experiments to see whether more evidence backs up a claim; one experiment is not enough.</td>
<td>Druger et al., 2004; Hunter et al., 2007</td>
</tr>
<tr>
<td>Seek opportunities to share their findings and communicate with others</td>
<td>Scientists present their work to others in the form of scientific posters, oral presentations, and written reports. Communication of their interpretations to the broader community is important, because scientists are working toward common goals.</td>
<td>Hurtado et al., 2009</td>
</tr>
</tbody>
</table>
mutation that changes one amino acid in the p53 protein sequence. Although much is known about p53 function, the specific functional defects of many mutant p53 proteins are uncharacterized. We used yeast as a model system because it is inexpensive to culture, grows rapidly, and is easily genetically manipulated. Using standard yeast genetics and an array of basic molecular and cell biology techniques, students can explore the phenotype of their p53 mutant over the course of a 10-wk quarter and come to an initial conclusion about the molecular nature of the defect in their mutant (Schärer, 1992; Mager and Winderickx, 2005; Figure 1).

**Course Goals: To Engage Students in a Research Experience to Encourage Scientific Thinking**

This course’s primary goal was to shift students from novice to expert in their thinking as scientists in the context of a scientific research project. To achieve this, we incorporated the following features that align with the defining features of CUREs (Auchincloss et al., 2014): 1) discovery and relevance: students explored one longitudinal research question in depth for 10 wk and come to an initial conclusion about the molecular nature of the defect in their mutant (Schärer, 1992; Mager and Winderickx, 2005; Figure 1).

2) collaboration: a high degree of collaboration required students to work with a partner on all aspects of the project, including experimentation, postlab assignments, and final presentations, and there were larger groups of students who shared data to achieve group conclusions; 3) iteration: multiple groups of students did the same experiments and compared data with one another; and 4) scientific practices: a) an emphasis on data interpretation and analysis and b) assessments that were representative of how scientists would evaluate one another, including a poster presentation and an oral presentation.

The course has a modular scalability in that five different p53 mutants were studied each term, and each lab section (10 students) studied the same five p53 mutants (one pair of students per mutant p53 allele). Even though the experimental procedures and the questions were mostly predetermined, the results and interpretations were unknown, creating a realistic research experience.

**Course Organization, Student Population, and Instructional Team**

Students enroll in this introductory lab course independent of an introductory biology lecture course. The course is intended for sophomore biology majors who are concurrently taking the introductory biology lecture series. All biology majors are required to take this course as their introductory biology lab course, and most premed students in nonbiology majors take this course to fulfill a medical school requirement; there were no other introductory biology lab course options available. Table 2 shows the demographic characteristics of students from Winter 2013, which is representative of the other terms.

The course comprised one 75-min lecture/discussion section and one 4-h lab each week. The lecture/discussion section was used to introduce new material, to give students an opportunity for guided practice with some of the concepts of the course, and to compare and contrast data from different mutants. Three exams were also given during these discussion sections.

The 4-h lab was primarily dedicated to conducting experiments. Lab partners were randomly assigned at the beginning of the course, and students worked with the same lab partners each week to investigate their specific mutants.

A PhD-level lecturer taught each discussion session to 20 students; this lecturer taught two adjacent lab sessions, each with 10 students, simultaneously. Additionally, a graduate student teaching assistant helped coordinate the lab sessions, so there was an overall teacher-to-student ratio

<table>
<thead>
<tr>
<th>Class year</th>
<th>Gender</th>
<th>Ethnicity</th>
<th>Major</th>
<th>Prior research experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sophomore 53%</td>
<td>Male 40.2%</td>
<td>White 39.3%</td>
<td>Biology 37.6%</td>
<td>Yes 60.7%</td>
</tr>
<tr>
<td>Junior 31.6%</td>
<td>Female 59.8%</td>
<td>Asian 41%</td>
<td>Human biology 37.6%</td>
<td>No 39.3%</td>
</tr>
<tr>
<td>Senior 15.4%</td>
<td>Black 9.4%</td>
<td>Other 8.5%</td>
<td>Engineering 16.2%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class year</th>
<th>Gender</th>
<th>Ethnicity</th>
<th>Major</th>
<th>Prior research experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sophomore 53%</td>
<td>Male 40.2%</td>
<td>White 39.3%</td>
<td>Biology 37.6%</td>
<td>Yes 60.7%</td>
</tr>
<tr>
<td>Junior 31.6%</td>
<td>Female 59.8%</td>
<td>Asian 41%</td>
<td>Human biology 37.6%</td>
<td>No 39.3%</td>
</tr>
<tr>
<td>Senior 15.4%</td>
<td>Black 9.4%</td>
<td>Other 8.5%</td>
<td>Engineering 16.2%</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Research questions explored by students in the course over a 10-wk period. Students work through each of these questions to determine the functional defect of their mutant p53. The experimental protocols for each experiment have been developed; thus, the authenticity of the course stems from the data analysis on mutant alleles that have not previously been characterized, as well as the progressive refinement of student-articulated hypotheses and conclusions.
of 1:10. Two tenure-track faculty members with expertise in yeast genetics and four PhD-level instructors whose primary teaching responsibility was this course were responsible for designing and implementing the curriculum.

Course Design Elements Designed to Promote Scientific Thinking

We integrated the following design elements throughout the course to promote student thinking like a scientist.

Query as a Way to Structure Student Thinking. QUERY is an acronym for “Question, Experiment, Results, and Your interpretation” (C. Anderson, unpublished observations). For each experiment students performed, they were asked to use the QUERY method to structure their thinking. This helped them to articulate the question they were trying to address, describe the experiment in detail, and differentiate results from their interpretation of the results. Specifically, students were asked about the question and experiment on each prelab assignment and the results and interpretation on each postlab assignment. Thus, we used QUERY as a way to scaffold the process of thinking like a scientist to introductory students. It required students to think about the “why” behind each experiment and distinguish the results from their interpretation of the results.

Hypothesis Testing and Making Predictions. Although the instructors knew the order of the experiments, which was essential for planning and providing reagents to such a large class, the students did not know the order of the experiments in advance. Thus, after students completed the first set of analyses to determine whether their mutant p53s had transactivation defects, we gave students the opportunity to brainstorm about what might be causing the defects and how they might be able to test their hypotheses. Students engaged in a brainstorming session during which they used inductive logic to ask what next set of experiments they should design to answer their overarching question: “What is wrong with your p53 mutant?” These brainstorming sessions provided opportunities for students to see the similarities and differences between experiments and experience the benefits of having multiple people working together to solve a problem. These sessions also prompted students to see the connections between individual experiments in answering the overall question (Figure 1).

Additionally, on each weekly postlab assignment, students were asked, based on data collected thus far, to develop hypotheses concerning possible molecular defects in their mutant p53s. This exercise could help students organize what they knew already and keep the big picture of the project in perspective. These activities were intended to help students see the project as one longitudinal project, even though they were completing a series of smaller experiments.

Data Interpretation. Each week’s postlab assignment focused on the data analysis and interpretation of experiments conducted in the lab. Partners worked together on the data analysis and interpretation, and each pair of partners submitted a single, collaboratively prepared, postlab assignment. Thus, the collaboration between them was realistic of how scientists in a research lab would collaborate.

Building in Iteration and Comparing Data. Within each lab room, each student pair worked on a different mutant, affording them ownership over their studies and independence in their interpretation of the data. However, students in different lab rooms examined the same set of five mutants. Three times during the 10-wk quarter, we held mutant group discussions, in which all the students working on the same mutant came together to compare their data and draw conclusions about their p53 mutants (Figure 2).

During one of these meetings, the group assigned to a specific mutant collectively designed an experiment, deciding how many different variables they wanted to test and weighing that against the benefit of having a higher number of replicates. These mutant group discussions allowed students to see, through the lens of their own p53 mutants, the inherent variability of biological data and, consequently, the importance of having multiple replicates of each experiment. When experiments did not yield an interpretable result, students engaged in a process of troubleshooting to determine what might have gone wrong. Differences between student results provided an opportunity to brainstorm possible sources of error and ways to make the data more reliable in the future; students also decided among themselves whether to include data points that could be considered outliers. Although instructors and teaching assistants facilitated these discussions, students were encouraged to lead the discussions, and student participation was integral.

We emphasized that the results of only one experiment are not enough to draw a conclusion; experiments must be repeated multiple times, which is a concept that students have been shown to have difficulty understanding (Brownell et al., 2014). Experiments must be repeated when they fail and when they work—one replicate is not enough to draw a conclusion in a biological experiment.

If an experiment did not yield interpretable results in a given week, students were expected to come in during another time before the next lab to repeat the experiment; this encouraged students to be diligent in their experimentation and also gave them a more realistic experience of what it...
situating student-generated data within what is known: accessing primary literature and p53 database. students read and discussed a primary scientific paper that was relevant to their investigation of p53 mutants. this exercise challenged students to think critically about published data and encouraged them to think about what other experiments they would need to do before their own work could be published (e.g., more replicates, additional experimental approaches to answer each question). the discussion also had students draw a connection between their work with yeast and the therapeutic implications in humans.

students also accessed an online p53 database to explore research that had previously been done on their p53 mutants and the number of human tumors that had been identified with their p53 mutations. this allowed them to recognize what elements of the research project were novel and how their experiments were situated within the growing body of scientific knowledge.

assignments representative of how scientists would present data. because student results were preliminary, we sought to mimic what scientists in a research lab would do with preliminary data: present them to their colleagues in lab meeting and present them in a poster venue. for the lab meeting presentation, all students who studied the same mutant in the lab section presented one set of representative data and their interpretations from all the experiments. as they presented, other students and instructors asked questions concerning data analysis and interpretation, often comparing them with other sets of data that had been presented. for the final poster presentation, each partner pair created a scientific poster, which they took turns presenting during the poster session, a large venue with more than 50 posters. students were required to visit other posters and compare other students’ results with their own results, completing graded worksheets that summarized their observations.

building community through pass/fail grading. the course was offered on a pass/fail basis, primarily because we wanted to create a community of collaborators rather than competitors. additionally, we felt that having students strive for a letter grade, and thus focus on the final answers and point values, would detract from them learning how to think like scientists. this perspective was confirmed by students enrolled in the course: more than 50% of students preferred that the course was offered pass/fail, and the most common reason cited was the desire to avoid having to worry about points on each assignment (unpublished data). it has also been shown that offering courses on a pass/fail basis rather than assigning a letter grade can lead to improvements in psychological well-being (bloodgood et al., 2009) and group cohesiveness (rohe et al., 2006); we were most interested in group cohesiveness, because we wanted the students to view themselves as collaborators.

accountability. because motivation can be a problem for pass/fail courses (gold, 1971), we implemented a policy that required students to earn an average of 70% on their postlab assignments and final posters. these standards encouraged students to take the course seriously and held them accountable for what they learned; each year, only one to two students failed the course due to low exam scores.

steps taken to implement a high-enrollment required cure. before this new course was implemented, the existing introductory biology lab course was a standard “cookbook” course in which the topics and organism studied changed every 2 wk and students worked through predetermined protocols to get a known “right” answer. the impetus for redesigning the course was a history of poor student evaluations for this course and a desire among faculty members to improve the course, in conjunction with national calls for biology lab course reform (nrc, 2003; aaas, 2011). because we were uncertain how students would react to this new course or whether this type of course would be possible for a large, introductory population of students, we decided to gradually scale up the course. the new lab course was introduced in 2 yr of pilot versions of the course. in winter 2010, a pilot version of the research-based course was taught to ~20 students who volunteered to take it rather than the older “cookbook” course. based on the positive reaction to the new course expressed in interviews and attitudinal and self-efficacy surveys (unpublished data), the course was scaled up to a larger group of students the next year. the course was taught in winter 2011 to ~40 students who were chosen at random (i.e., nonvolunteers) into either the pilot course or the existing cookbook lab course, and evaluation of the pilot course was largely positive (unpublished data). in winter 2012, the new course was implemented to 250+ students, and four phd-level instructors taught different sections of the course. increased collaboration between students within the lab sections and in mutant groups was added to the course, as was a poster session at the end of the course. the curriculum that we are assessing is the most recent curriculum, which has now been taught in approximately the same way for five quarters (winter 2012, fall 2012, winter 2013, fall 2013, and winter 2014) to a large-enrollment population of nonvolunteer students, giving us confidence that implementing this type of course in a large introductory setting as a required course is possible.

evaluation

methods. to assess the course, we used an approach of analyzing pre- and postcourse open-ended written responses and course exams. because time for assessment was limited in the 10-wk quarter, we have collected different types of data from the two times the course has been offered as a required component of the introductory curriculum (fall 2012 and winter 2013). students were given pre- and postcourse surveys that included open-ended questions about their conception of what it meant to think like a scientist. question prompts were designed through a series of think-aloud interviews with students. student responses to open-ended questions were coded using a combination of content analysis and
grounded theory (Glaser, 1978; Glaser et al., 1968) to identify themes, from which specific categories were chosen. Two independent raters scored a subset of student responses, and they came to a consensus when they disagreed. The frequency of each student response was calculated for each category. Student responses could include more than one idea, so responses do not sum to 100%. All students enrolled in the course completed the surveys, and a random subset of student responses was used in the analysis; lab sections were chosen at random, and all student responses in those sections were included in the analysis. Post hoc, we made sure that student responses were included for each instructor and time slot (day and time) to minimize the bias that may be caused by only examining student responses from one instructor or from one section time. Data used are from Fall 2012 \( (n = 60) \) and Winter 2013 \( (n = 117) \). Statistical analysis was done using paired \( t \) tests \( (p < 0.05) \).

Additionally, open-ended questions on the postcourse surveys asked students whether their thinking like a scientist had changed as a result of the course and, if so, how their own thinking like a scientist had changed. Data used are from Fall 2012 \( (n = 60) \). Students were also asked specific Likert-scale questions about what components of the course were important for their understanding of thinking like a scientist. The average Likert score for each question and the SD were calculated. Data used are from Winter 2013 \( (n = 117) \).

Students’ ability to design experiments and interpret data were measured by assessing their scores on three exams. We determined the composition of questions focused on data analysis by having two independent raters review exam questions. We found that 50% of the total points on the exams asked students to read graphs or interpret figures and tables and 21% of the total points on the exams were questions that verbally described experimental results or conditions, so 71% of the exam points were specifically directed at eliciting student understanding of data analysis and interpretation. Thus, the composition of the exam questions was predominantly data analysis or interpretation questions, giving us confidence that this could be a measure of student ability to analyze and interpret data.

We characterized the cognitive level of each question on the exams (Crowe et al., 2008) and calculated a weighted Bloom average for each exam adapted from Freeman and colleagues (2011). We also calculated a difficulty score for each question, using a rubric that was developed by the course instructors (Table 3). To do this, we scrambled all the exam questions so the rater would not know which question came from which exam. We had two former instructors for this course blindly score the questions according to Bloom’s level and difficulty. They achieved greater than 80% interrater reliability on the Bloom’s level and 100% interrater reliability on the difficulty score. These former instructors did not teach the course in the term that used these particular exams, so they did not have knowledge about specific questions on specific exams. Additionally, these former instructors had an expert understanding of the course content and familiarity with what was explicitly covered in the course, so they knew how the data were typically presented to students in the course and what specific experiments students conducted. Scores were compared and discussed until consensus was achieved. Data used are from Winter 2013.

While a common method of assessment in biology education is a pre–post test format, we chose not to give students a pretest focused on data analysis at the beginning of the course, because there was too much content-dependent information that students would not have known (e.g., how to interpret specific assays). We also chose not to give students a pre–post test on their content knowledge, because that was not one of our course goals. We were using yeast genetics to explore p53 and cancer as a model system; students had to learn certain details about the system to be able to pose hypotheses and analyze data, but we focused on content information only to the extent that it was necessary for them to learn the process skills. It would be interesting to see how students perform on a pre–post test of their ability to interpret data in a content-independent way, but at the time of our study, no such assessment tool existed.

### Results

**Finding 1: Students Show a More Expert-Like Conception of What It Means to Think Like a Scientist at the End of the Course and Perceive That Their Own Thinking Has Changed.** We found that student understanding of what it means to think like a scientist became significantly more nuanced and similar to expert scientists’ thinking when comparing their postcourse and precourse answers to the question: “What do you think it means to think like a scientist?” (Table 4). Specifically, students at the beginning of the course mentioned being curious, being critical or logical, developing hypotheses, or using the scientific method. However, at the end of the course, the responses were more grounded in their lab experience, including a focus on collaboration and data analysis. Specifically, students at the end of the course mentioned needing to be skeptical of data, the need to repeat experiments, how scientists can learn from failed experiments, and how there are multiple ways to approach a problem (Table 4).

When we asked students at the end of the course whether their own thinking like a scientist had changed during their investigation of mutant p53, we found that 100% of students \( (n = 60) \) thought that their thinking like a scientist had changed (Table 5). In response to an open-ended question of how their thinking like a scientist changed, 83% of student responses could be classified as thinking scientifically,

<table>
<thead>
<tr>
<th>Table 3. Difficulty rubric for exam questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy (1) Definitions/explanations of what was previously presented in lab (e.g., purpose of a particular step of an experiment or fact about cancer)</td>
</tr>
<tr>
<td>Medium (2) Students need to apply their knowledge to a situation that they experienced in lab or analyze the results of one graph/figure in the same way they analyzed it in lab (e.g., students have to predict what went wrong when given an experimental result)</td>
</tr>
<tr>
<td>Difficult: complex data or near transfer (3) Students need to apply their knowledge to either a complex set of data (more than one figure at once) or to data presented in a novel way (e.g., students interpreting an unfamiliar graphical representation of data).</td>
</tr>
</tbody>
</table>
### Table 4. Student responses to the open-ended question “What does it mean to think like a scientist?”

<table>
<thead>
<tr>
<th>Theme</th>
<th>Percentage of responses categorized under this theme</th>
<th>Example student responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Involved collaboration</td>
<td>Precourse: 0 Postcourse: 20.5*</td>
<td>“This quarter has taught me the importance of working collaboratively with others in order to more fully understand the topic of research.” “Be willing to collaborate.”</td>
</tr>
<tr>
<td>Requires analyzing and interpreting data</td>
<td>Precourse: 6.0 Postcourse: 31.6*</td>
<td>“Thinking like a scientist requires lots of analyzing of data and asking so what? Why is this? What is next?” “One has to analyze the results and try to interpret them then draw up other experiments that can confirm the results.”</td>
</tr>
<tr>
<td>Being skeptical of data</td>
<td>Precourse: 0 Postcourse: 18.8*</td>
<td>“It also means to be skeptical and critical of data, and to never trust just one set of data but try to continuously strive for accurate and less variable results.” “To not be stubborn and ignore results that contradict with your hypothesis.”</td>
</tr>
<tr>
<td>Need to repeat experiments</td>
<td>Precourse: 2.6 Postcourse: 12.8*</td>
<td>“Test and retest.” “Additionally, I have learned the importance of repeating experiments and testing hypotheses in a variety of ways in order to gain more significant data.”</td>
</tr>
<tr>
<td>Learn from mistakes/failed experiments</td>
<td>Precourse: 1.7 Postcourse: 9.4*</td>
<td>“Troubleshooting experiments that don’t go as planned, i.e., designing experiments to figure why the original experiment wasn’t working.” “Identify why errors occurred.”</td>
</tr>
<tr>
<td>QUERY as a way to organize thinking for each experiment</td>
<td>Precourse: 0 Postcourse: 15.4*</td>
<td>“It means asking a question, designing an informed hypothesis based on background knowledge, creating an experiment to test this question and interpreting the results. (QUERY)” “QUERY is a huge part in being able to think like a scientist.”</td>
</tr>
<tr>
<td>Using multiple approaches to answer a question; many ways of thinking</td>
<td>Precourse: 9.4 Postcourse: 17.9*</td>
<td>“As a scientist, you want to approach a topic or research points from multiple angles. There may be one experiment that shows X, but you always want to verify that result with other experiments. As a scientist, you want to realize that experiments have limitations and by having multiple experiments to support one another, you can synthesize a conclusion from all the data.” “Do a series of experiments to try and answer it while using both qualitative and quantitative methods.”</td>
</tr>
<tr>
<td>Critical, logical thinking</td>
<td>Precourse: 25.6 Postcourse: 21.4</td>
<td>“Thinking like a scientist means thinking critically and thinking through all possibilities in order to best devise controls and alternate experiments.” “You have to make a conscious effort to not jump to conclusions you want to be true and only take what the data gives [sic] you.”</td>
</tr>
<tr>
<td>Developing hypotheses</td>
<td>Precourse: 21.4 Postcourse: 29.1*</td>
<td>“You try to learn more about the world around you by creating experiments and testing hypothesis [sic].” “Form educated hypotheses.”</td>
</tr>
<tr>
<td>Using the scientific method</td>
<td>Precourse: 16.2 Postcourse: 1.7*</td>
<td>“Think like a scientist means to constantly employ the scientific method—observe, hypothesize, experiment, conclude—in all aspects of life in order to reach better understandings of different topics.” “Thinking like a scientist means to think in a logical and organized method, in particular adhering to the scientific method. In such a method, research begins with an observation, hypothesis, question prediction, followed by the research and data and results.”</td>
</tr>
<tr>
<td>Generalized vague statement about being curious</td>
<td>Precourse: 23.9 Postcourse: 16.2*</td>
<td>“To be curious about how and why things work and to then pursue those curiosities with experimentation.” “To be inquisitive. A desire to learn how/why things work and how things break down/what causes them to malfunction.”</td>
</tr>
</tbody>
</table>

*p < 0.05, paired t tests.

Data from Winter 2013 (n = 117).
as follows: 1) collaboration is important (20%), 2) be skeptical of data and make only tentative conclusions (38%), 3) repeat experiments (27%), 4) learn from making mistakes/failed experiments (13.3%), 5) use multiple approaches to answer a problem (13.3%), and 6) controls are important. Thus, not only did students have a clearer idea of what it meant to think like a scientist, they felt as though they had acquired these skills as a result of taking this course.

Finding 2: Students Indicated That Specific Aspects of the Course Focused on Data Analysis and Collaboration, Including the Mutant Group Discussions, Were the Most Useful for Their Learning How to Think Like a Scientist. At the end of the course, students were asked in an open-ended question what specific aspects of the course were most useful in helping them to think like a scientist. The most frequent themes that emerged after analyzing the data that captured 92% of student responses included (Table 6): 1) mutant group discussions (27%), 2) data analysis aspect of postlab assignments (26%), 3) performing different experiments on one longitudinal question (24%), 4) predicting results and comparing actual results with predicted ones (15%), 5) troubleshooting failed experiments (8%), and 6) collaborating with other students in the class (8%).

In addition, we gave students a list of some of the design elements of the course and asked them to rate these items on a five-point Likert scale on how useful a particular component of the course was for improving their thinking like a scientist (1, not at all useful; 2, slightly useful; 3, moderately useful; 4, very useful; and 5, extremely useful). Once again, we found that components of the course that related to collaboration and data analysis were most highly rated (Table 7), with the one exception being “your instructor’s teaching.” However, the majority of instruction that occurred in the lab was focused on data analysis, so it could be that students were referencing the instructor teaching them this skill.

An interesting comparison between the open-ended data and the Likert-scale data is that mutant group discussions was the most frequent response for the open-ended question but was only rated in between moderately useful and very useful on the Likert scale. Additionally, the SD is large, indicating that this aspect of the course may be polarizing for students or that perhaps the quality of the different mutant...
ment in the mutant groups was captured by the following responses:

- "Sometimes we failed to reach a conclusion, since often we all had conflicting results."
- "[Mutant groups] were drawn out and inconclusive."
- "The greatest area for improvement for the mutant discussion group is reaching a more firm and assertive group conclusion."

These students missed the idea that purpose of the mutant group discussions was to illustrate the inherent variability of scientific data and the inappropriateness of drawing conclusions based on one replicate of one experiment. Students wanted an answer, the “right” answer, and were frustrated when effort was spent and no conclusion could be reached.

However, the other group of students (~25% of responses) seemed to understand that the data they had collected with their own hands were one set out of a larger collaborative data set that they would use to determine the functional defect of their mutant p53; thus, the mutant group discussions became paramount for these students’ understanding.
of the significance of repetition of experiments and in statistical analysis of results, and their responses captured this:

“Being able to compare your data to better understand what next steps to take and what was actually happening with our p53 mutation.”

“The opportunity to be away from the lab and have the time to sit and actually process and discuss what we’re doing in class. Also the opportunity to talk to other mutant groups and think about what could have gone wrong, what are the general trends etc. was really great in actually understanding the lab content. I think this was the best learning experience of the entire lab.”

Finding 3: Students Showed Improvement in Their Ability to Analyze and Interpret Data. We designed three exams to test student ability to analyze and interpret data, including understanding the significance of biological variation, repetition of experiments, and how to analyze the data they had obtained during their investigations into the functional defects of p53. Some of the questions were based on specific experiments they had conducted, but others required students to analyze data in a novel way or to transfer their knowledge to a novel scenario (see Supplemental Material for sample questions).

Average student scores on these exams remained constant over the term; students’ average scores were 87.1% on exam 1, 86.6% on exam 2, and 88.2% on exam 3 (Table 9). Student performance on exams was not curved or normalized in any way. While the total score remained constant, the Bloom’s level of the average exam question increased significantly over the term (Kruskal-Wallis $\chi^2 = 13.7$, $df = 2$, $p = 0.001$). Average weighted Bloom levels were 2.69, 2.91, and 3.61 for exams 1, 2, and 3, respectively. Additionally, we designed a “difficulty scale” based on whether students were giving an explanation of what was already presented in the lab course, were examining data in the same way they had been presented the data in class, or were faced with a more complex set of data presented in a novel way. Based on this difficulty scale, exam 1 was 1.61, exam 2 was 1.82, and exam 3 was 2.41. Thus, the exams had a significant increase in the difficulty of the average exam question (Kruskal-Wallis $\chi^2 = 10.35$, $df = 2$, $p = 0.006$).

DISCUSSION

In this study, we present an innovative curriculum that allows students to experience research in the context of an introductory lab course. One of the strongest aspects of this course is that it is required of all biology majors, thus giving all biology majors exposure to research at the undergraduate level. Many CUREs have only been offered to a small number of volunteers, which limits the ability to generalize conclusions drawn from the evaluation of these smaller courses. With the scale-up of this course, every biology major at this institution engages in a research project as part of his or her undergraduate curriculum. Additionally, it makes scientific research more inclusive by limiting the number of hurdles students have to jump through to engage in research and by removing the selection process that faculty members use to pick students for individual apprenticeships (Bangera and Brownell, 2014).

Gains in Student Conceptions of What It Means to Think Like a Scientist

The theme of thinking like a scientist has appeared in other studies of traditional independent research apprenticeships, although in contrast to these other studies, our definition did not include deeper conceptual knowledge about the topic (Seymour et al., 2004; Hunter et al., 2007). We limited our

Table 7. Aspects of the course that were most useful for improving your thinking like a scientista

<table>
<thead>
<tr>
<th>Component of course</th>
<th>Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analyzing your own data</td>
<td>4.35 (0.68)</td>
</tr>
<tr>
<td>Working with a partner on all aspects of the project</td>
<td>4.30 (0.87)</td>
</tr>
<tr>
<td>Your instructor’s teaching</td>
<td>4.23 (0.85)</td>
</tr>
<tr>
<td>Comparing your data with data from other groups working on the same mutant in your lab section</td>
<td>4.16 (0.81)</td>
</tr>
<tr>
<td>Completing postlab assignments</td>
<td>4.12 (0.77)</td>
</tr>
<tr>
<td>Creating a poster</td>
<td>3.45 (1.16)</td>
</tr>
<tr>
<td>Mutant group discussions</td>
<td>3.42 (1.24)</td>
</tr>
<tr>
<td>Reading a primary scientific paper</td>
<td>3.37 (1.14)</td>
</tr>
<tr>
<td>Using the QUERY method</td>
<td>3.32 (1.13)</td>
</tr>
<tr>
<td>Comparing your data with data from other groups working on a different mutant in your lab section</td>
<td>3.27 (1.19)</td>
</tr>
<tr>
<td>Designing your own experiment</td>
<td>3.18 (1.13)</td>
</tr>
<tr>
<td>Repeating experiments that did not work the first time</td>
<td>3.12 (1.11)</td>
</tr>
<tr>
<td>Working through the handouts during lab</td>
<td>3.07 (1.03)</td>
</tr>
<tr>
<td>Course exams</td>
<td>3.05 (0.91)</td>
</tr>
</tbody>
</table>

a Students evaluated this question on a closed-ended Likert scale (1, not at all useful; 2, slightly useful; 3, moderately useful; 4, very useful; 5, extremely useful). Data from Winter 2013 ($n = 117$).
an understanding that it is a process that is responsive to experimental results and constantly evolving in the face of new data, new techniques, and new ideas. This emphasis on data was echoed when students were asked about the specific aspects of the course that contributed to an improvement in their thinking like a scientist: data-centered collaborative aspects of the course were highlighted. Students focused on mutant group discussions and postlabs as sources of scientific thinking. Building multiple opportunities for students to think critically about data, including having them repeat experiments, helped students grasp the variability that is inherent to experimental science and the need for multiple, independent observations to build convincing support for a hypothesis.

However, we saw that many students still struggled in the mutant groups with not being able to find a “right” answer. This may indicate that we need to integrate more opportunities for mutant group discussions into the course and more explicit discussion of the importance of biological replicates in order to come to a conclusion.

Table 8. Single greatest strength of the mutant group discussions*

<table>
<thead>
<tr>
<th>Theme</th>
<th>Percentage of responses categorized under this theme</th>
<th>Example student responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confirming that “their data” were similar to others by comparing results</td>
<td>54.7</td>
<td>“The single greatest strength of the mutant group discussion was to have the opportunity to compare our data with that of the other mutant groups.”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Being able to see that our results were not completely askew.”</td>
</tr>
<tr>
<td>Achieving consensus about functional defect of mutant p53</td>
<td>24.8</td>
<td>“Verify/compare results to the same experiment to simulate conducting multiple trials.”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“It was very beneficial to see some consensus or lack of consensus between teams; it put everything more in perspective.”</td>
</tr>
<tr>
<td>Discussing possible sources of error in experiments</td>
<td>13.7</td>
<td>“Discussing the results each group got and which were the possible correct results. Determining reasons for any outliers.”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“An opportunity to discuss possible sources of variation and the strength of our controls.”</td>
</tr>
<tr>
<td>Representative of an authentic lab discussion about data</td>
<td>6.0</td>
<td>“A preview of what collaboration in a lab is actually like (people sometimes get completely different results and those differences need to be reconciled).”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Collaborating was always interesting and made our results feel much more real.”</td>
</tr>
</tbody>
</table>

*Data from Winter 2013 (n = 117).

In pilot studies, we saw that students’ confidence in their ability to analyze data improved as a result of this course and through exam analysis (unpublished data). Confidence

Table 9. Student ability to analyze data improved as the exams got more difficult and their exam scores stayed constant

<table>
<thead>
<tr>
<th>Average student score</th>
<th>Weighted average Bloom score for each exam</th>
<th>Weighted average difficulty score for each exam</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exam 1 87.1%</td>
<td>2.69</td>
<td>1.61</td>
</tr>
<tr>
<td>Exam 2 86.6%</td>
<td>2.91</td>
<td>1.82</td>
</tr>
<tr>
<td>Exam 3 88.2%</td>
<td>3.61</td>
<td>2.41</td>
</tr>
</tbody>
</table>

*Data from Winter 2013 (n = 117).
As the difficulty of exams increased, students’ average scores remain constant. This is particularly encouraging, because these exams were not high-stakes exams. We wanted to establish a culture of collaboration rather than competition, so we chose to have the course be ungraded and taken only as pass/fail. Students were required to obtain 70% on the average of the three exams but were not rewarded for performing better than that. Despite the potential lack of motivation, we still saw student improvement in the ability to analyze data. This is important, because most assessments of CUREs rely on student self-reporting of their abilities rather than one measurement of their actual ability; we were able to document learning gains, and others interested in assessing CUREs could use a similar method.

Advantages of a Large-Enrollment Course
Finally, rather than a large course enrollment being an inconvenience, the specific format of the course actually benefits from having large numbers of students working on the same mutant alleles and sharing data. More students meant a larger number of replicates and higher probability of seeing a conclusive pattern in the data. We acknowledge that our enrollment of ∼200–300 students, while large by the standards of our institution, is not large by the standards of others that need to cater to more than 1000 introductory-level students. However, we believe that this course could be scaled up further and are interested in working with others to help adopt this course, or similar courses, at their institutions.

LIMITATIONS OF INTERPRETATIONS
Teacher Effect May Be Acting Independently of the Curriculum
This course was taught by PhD-level instructors, which likely contributes to an “instructor effect” that has an impact distinct from the curriculum itself on student learning. Our goal for this study was to examine the course curriculum rather than an individual instructor’s teaching, but the two by their very nature are entwined. It could be possible that the gains we saw are due to the synergistic effects of the research-based curriculum and having expert scientists as instructors. We specifically decided to have PhD-level instructors, because we wanted instructors with expertise in scientific thinking, but we hope that it is possible to replicate the same quality of high-level instruction by using more advanced graduate students or postdoctoral fellows.

It is also important to note that instructors all had very different teaching styles, ranging from a more traditional lecturing to more interactive active-learning teaching, and diverse training in biological sciences. While we do not see notable differences among instructors on most of the assessments, leading us to conclude that the major source of impact for the course is derived from the common curriculum, there was one instance in which particular instructors had significant differences. In Fall 2012, for two of the instructors, more than 90% of the student responses could be categorized as “thinking like a scientist,” whereas this was only 50% for one of the other instructors. For others interested in developing this type of course and having the course taught by instructors with diverse teaching experience and/or styles, it will be important to determine whether instructors with limited teaching experience (e.g., first-year graduate students) can adequately facilitate this type of course to achieve the desired learning gains.

Student Population May Limit Generalizations
Another critical caveat arising from our evaluation of the benefits of this type of course for an introductory population of students is that the student population at Stanford includes a high proportion of high achievers. Even though this is the first biology lab experience for them in college, a significant percentage of our students have had previous research experience. Although we acknowledge that this is likely perceived as a limitation to generalizing our results to all universities, there are students in this cohort who have no prior research experience and a weak biology background. Furthermore, unlike many CUREs, students are required to take this course, so there is no self-selection bias.

Although our population is high achieving, it is considerably ethnically diverse. White students make up less than 40% of the class, with almost 20% of the students coming from historically underrepresented minority populations. Additionally, ∼15% of Stanford undergraduates are the first in their family to attend a 4-yr college.

Cost of Running This Type of Course
Although we encourage others to adopt a similar curriculum, this course does have an associated cost that may be a barrier for many institutions. We were fortunate to receive several grants to ensure ongoing commitment to the course. In addition to start-up costs (e.g., equipment), there are high costs for experimental materials and supplies (e.g., antibodies, the DNA-binding assay kits). Although there are ways to decrease the cost of running this course by not doing particular experiments, the cost of running a molecular course is inherently higher than other types of biological courses. As budgets become tighter, this does present a logistical challenge to those interested in developing this type of course. As a contrast to this high-cost course, we have developed another research-based lab course that integrates teaching and research; it focuses on ecological relationships and is a low-cost alternative that also gives students a research experience in the context of an introductory biology lab course. For an overview and assessment of that course, see Brownell et al. (2012) and Kloser et al. (2013).

CONCLUSION
We encourage others to consider developing this type of high-enrollment course; it allows students the opportunity to think about larger data sets than would be possible from a small 10–20 person class. Additionally, the opportunities for collaboration are extensive. However, a larger class is logistically more challenging. We encourage others to start small and then build from there; our first pilot year we had
only 20 students and few of the experiments worked. In the second
year, more of the experiments worked, but kinks
still needed to be worked out. It took many iterations of
the course to work out the logistical challenges associated
with this type of course, and even after seven iterations, not
all of the experiments work every time. The most frequent
complaints from students the first two times the course was
offered centered on course logistics rather than on content,
curriculum, or instruction. While logistics may seem insig-
nificant and often are less of a problem in a lecture-based
course, they can be significant barriers to achieving course
goals in a lab course.

This course is a way to expose all students, not a select
few, to the joys and challenges of research, and it gives them
a more realistic impression of how research is done. We urge
others to consider adopting this type of course-embedded
research experience as a way to make research more acces-
sible to a larger population of students who likely would
not pursue independent research on their own (Bangera and
Brownell, 2014). This course, and others like it, can be in-
strumental for providing all students with the opportunity
to learn how to do science.

ACKNOWLEDGMENTS

Funding for this course came from a National Science Foundation
TUES (0941984), a Haagland Award for Innovations in Undergrad-
uate Teaching, an HHMI education grant, and a generous gift from
a private donor. We thank the 700+ students who have participat-
ed in this course, in particular the students who took the course
during the initial pilot phases, and the graduate student teaching
assistants who helped teach the course, particularly Janet Wegner
and Dan Van de Mark. Charles Anderson (now at Penn State Uni-
versity) first coined “QUERY” as an educational concept, and it has
now been incorporated into the course. We are especially grateful
to Nicole Bradon for her technical assistance in preparing reagents
and equipment for the weekly student laboratories. We also ac-
knowledge the support of colleagues in the Department of Biology,
including Shyamala Malladi, Matt Knope, Waheeda Khalfan, Tad
Fukami, and the leadership of our former department chairman,
Bob Simoni. Additionally, Rich Shavelson and Matt Kloser at the
Stanford Graduate School of Education were instrumental in early
discussions of course assessment.

REFERENCES

American Association for the Advancement of Science (2011). Vision
and Change in Undergraduate Biology Education: A Call to Action,
Washington, DC.

Auchincloss LC, Laursen SL, Branchaw JL, Eagan K, Graham M,
(2014). Assessment of course-based undergraduate research experi-

Bangera G, Brownell SE (2014). Course-based undergraduate re-
search experiences can make scientific research more inclusive. CBE

Bloodgood RA, Short JG, Jackson JM, Martindale JR (2009). A
change to pass/fail grading in the first two years at one medical
school results in improved psychological well-being. Acad Med 84,
655–662.

for measuring the effectiveness of course-based undergraduate re-
search experiences in undergraduate biology. Studies High Educ 40,
525–544.

Brownell SE, Kloser MJ, Fukami T, Shavelson R (2012). Undergradu-
ate biology lab courses: comparing the impact of traditionally based
“cookbook” and authentic research-based courses on student lab

matters: volunteer bias, small sample size, and the value of com-
parison groups in the assessment of research-based undergraduate

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

Coil D, Wenderoth MP, Cunningham M, Dirks C (2010). Teaching
the process of science: faculty perceptions and an effective method-

undergraduate research experiences: an agenda for future research

menting Bloom’s taxonomy to enhance student learning in biology.
CBE Life Sci Educ 7, 368–381.

Druger M, Siebert ED, Crow LW (eds.) (2004). Teaching Tips: Innova-
tions in Undergraduate Science Instruction, Arlington, VA: NSTA Press.

World 27, 48.

Freeman S, Haak D, Wenderoth MP (2011). Increased course struc-
ture improves performance in introductory biology. CBE Life Sci
Educ 10, 175–186.

human MSH2 missense mutations using yeast as a model sys-
tem: a laboratory course in molecular biology. Cell Biol Educ 3,
31–48.

Glaser BG (1978). Theoretical Sensitivity: Advances in the Meth-
Press.

Glaser BG, Strauss AL, Strutzel E (1968). The discovery of ground-

Gold RM (1971). Academic achievement declines under pass-fail

Handelsman J, Ebert-May D, Beichner R, Bruns P, Chang A, DeHaan


Hunter AB, Laursen SL, Seymour E (2007). Becoming a scientist: the
role of undergraduate research in students’ cognitive, personal, and

Scientist: Investigating the Outcomes of Introductory Science and Math
Courses, Atlanta, GA: Association of Institutional Researchers Forum.

Jordan TC, Burnett SH, Carson S, Caruso SM, Class K, DeJong RJ,
implementable research course in phage discovery and genomics
for first-year undergraduate students. mBio 5, e01051.

teaching and research in undergraduate biology laboratory educa-

research-based ecology lab course: a study of nonvolunteer achieve-
ment, self-confidence, and perception of lab course purpose. J Coll
Sci Teach 42, 90–99.

Levine AJ (1997). p53, the cellular gatekeeper for growth and divi-


PORTAAL: A Classroom Observation Tool Assessing Evidence-Based Teaching Practices for Active Learning in Large Science, Technology, Engineering, and Mathematics Classes

Sarah L. Eddy, Mercedes Converse, and Mary Pat Wenderoth

Department of Biology, University of Washington, Seattle, WA 98195-1800

Submitted June 4, 2014; Revised January 8, 2015; Accepted January 12, 2015
Monitoring Editor: Jeff Schinske

There is extensive evidence that active learning works better than a completely passive lecture. Despite this evidence, adoption of these evidence-based teaching practices remains low. In this paper, we offer one tool to help faculty members implement active learning. This tool identifies 21 readily implemented elements that have been shown to increase student outcomes related to achievement, logic development, or other relevant learning goals with college-age students. Thus, this tool both clarifies the research-supported elements of best practices for instructor implementation of active learning in the classroom setting and measures instructors’ alignment with these practices. We describe how we reviewed the discipline-based education research literature to identify best practices in active learning for adult learners in the classroom and used these results to develop an observation tool (Practical Observation Rubric To Assess Active Learning, or PORTAAL) that documents the extent to which instructors incorporate these practices into their classrooms. We then use PORTAAL to explore the classroom practices of 25 introductory biology instructors who employ some form of active learning. Overall, PORTAAL documents how well aligned classrooms are with research-supported best practices for active learning and provides specific feedback and guidance to instructors to allow them to identify what they do well and what could be improved.

Compared with traditional “passive” lecture, active-learning methods on average improve student achievement in college science, technology, engineering, and mathematics (STEM) courses (Freeman et al., 2014). Unfortunately, the quantity and quality of evidence supporting active-learning methods has not increased faculty and instructor adoption rates (Fraser et al., 2014). Thus, although the development of new and optimized classroom interventions continues to be important, many national agencies concerned with undergraduate education have broadened their efforts to include a call for the development of strategies that encourage the broader adoption of these research-based teaching methods at the college level (President’s Council of Advisors on Science and Technology, 2012; National Science Foundation [NSF], 2013).

Strategies developed to encourage faculty adoption of active-learning practices need to acknowledge the realities of faculty and instructor life and the many potential barriers to adoption identified in the literature. At the institutional level, these barriers include a reward system that can lead faculty members to devote less time and effort to teaching (Lee, 2000) and limited institutional effort to train graduate students or faculty members on teaching methods (Cole, 1982; Weimer, 1990). At the individual level, faculty members may not identify as teachers and therefore fail to put
the same effort into their teaching as they do their research, may not recognize that the teaching strategies they use are not as effective as other strategies, or may not recognize that comfort and familiarity often dictate their choices of teaching method (Cole, 1982; Bouwma-Gearhart, 2012; Brownell and Tanner, 2012).

Even when faculty members are interested in learning new teaching practices, there remain multiple challenges to the effective implementation of those practices. First, there is a lack of clarity as to what active learning is, and this lack of clarity can lead to lack of fidelity of implementation of the teaching technique (O’Donnell, 2008; Turpen and Finkelstein, 2009; Borrego et al., 2013). For example, when asked to define active learning, faculty members in one study offered a range of answers, including simply “using clickers” or “group work” to “emphasizing higher-order skills” (Freeman et al., 2014). Many of these responses seem to conflate the tools used (clickers) to facilitate active learning with the actual methodology of active learning. In truth, both active-learning teaching methods and student learning are complex processes that do not have a single agreed-upon definition and can involve many components both inside and outside the classroom (Figure 1). This complexity, in and of itself, is a second barrier: changes in teaching practice can feel overwhelming because of the number of aspects that must be considered, and this can lead to paralysis and inaction (Kreber and Cranton, 2000). Finally, faculty members and instructors may not be familiar with the findings from the education research literature. With so many conflicting demands on their time, few faculty members have the time to immerse themselves in the education research literature. Furthermore, the education research literature frequently does not provide sufficient detail for the proper implementation of the educational innovations presented (Borrego et al., 2013).

We propose that the education research community interested in faculty change could better influence adoption of research-based best practices if we developed tools to efficiently communicate education research to instructors. Whatever these tools may be, they should include a clear description of the critical components of each teaching intervention such that a novice could implement them. Such an effort would provide faculty members with a clear set of instructions that would help them more readily implement teaching innovations with a higher degree of fidelity and, thus, possibly encourage the broader adoption of these research-based best practices for active learning. This paper is one attempt to create a tool that both clarifies the research-supported elements of best practices for instructor implementation of active learning in the classroom setting and helps instructors measure their alignment with these practices. This tool, Practical Observation Rubric To Assess Active Learning (PORTAAL), cannot address all the components of the complex learning environment (Figure 1) but can provide an accessible and informative entry point for implementing active learning in the classroom.

INTRODUCING PORTAAL: A PRACTICAL OBSERVATION RUBRIC TO ASSESS ACTIVE LEARNING

PORTAAL is intended to provide easy-to-implement, research-supported recommendations to STEM instructors trying to move from instructor-centered to more active learning–based instruction. For this paper, we operationally define active learning as any time students are actively working on problems or questions in class. We realize this is not an all-encompassing definition, but it is appropriate for the one area we are focusing on: behaviors in the classroom. Three goals guided the development of the tool:

Goal 1. PORTAAL is supported by literature: We identified dimensions of best practices from the education research literature for effective implementation of active learning. These practices are independent of particular active-learning methods (POGIL, case studies, etc.).

Goal 2. PORTAAL is easy to learn: We translated dimensions of best practices into elements that are observable and quantifiable in an active-learning classroom, making the tool quick and relatively easy to learn.

Goal 3. PORTAAL is validated and has high interrater reliability: The tool focuses on elements that do not require deep pedagogical or content expertise in a particular field to assess, making it possible for raters with a wide range of backgrounds to reliably use the tool.

Figure 1. PORTAAL captures one aspect of active learning: how the instructor structures the in-class experience. Active learning is a multifaceted practice that involves inputs from the instructor and students as well as events in and outside class. All these inputs influence the ultimate outcome of student learning.
Although there are many classroom observation tools available (American Association for the Advancement of Science, 2013), none of these tools was explicitly designed to capture evidence of the implementation of research-supported critical elements for active learning in the classroom. Thus, in this paper, we will present our classroom observation tool, PORTAAL, and preliminary data demonstrating its effectiveness. Part 1 demonstrates how we used the three goals described above to develop our observation tool that is evidence-based, easy to use, and reliable. In Part 2, we demonstrate how we used PORTAAL to explore how active learning is being implemented in the classroom by documenting the teaching practices observed in a range of introductory biology classrooms at one university. Although PORTAAL was tested with biology courses, this instrument should work in any large STEM classroom, as the best practices are independent of content. Part 3 offers recommendations for how PORTAAL could be used by instructors, departments, and researchers to better align classroom teaching practices with research-based best practices for active learning.

**PART 1: PORTAAL DEVELOPMENT AND TESTING**

**Goals 1 and 2: Identifying Research-Based Best Practices for Large Classrooms (Literature Review)**

The validity of PORTAAL is based in the research literature. Each of the 21 elements is supported by at least one published peer-reviewed article that demonstrates its impact on a relevant student outcome.

Elements for inclusion in PORTAAL were gathered from articles and reviews published from 2008 to 2013 focused on classroom implementations or other relevant research on adult learning (i.e., lab studies) that documented changes in one of four types of outcomes: 1) improvement in student achievement on either formative (e.g., clicker questions, practice exams) or summative (e.g., exams) assessments, 2) improvement in student in-class conversations in terms of scientific argumentation and participation, 3) improvement in student self-reported learning when the survey used has been shown to predict student achievement, and 4) improvement of other measures related to student performance or logic development. Relevant other measures could include an increase in students’ self-reported belief that logic is important for learning biology or an increase in the number of students who participate in activities. These articles came from discipline-based education research (DBER) and cognitive, educational, and social psychology fields. Thus, the methods used to measure student outcomes vary widely. Even within a single outcome, the way the outcome was measured varied from study to study. For example, student achievement was measured in a range of ways: some studies measured student clicker question responses in a classroom, others looked at exam scores, and still others took measurements in psychology laboratory settings. The differences in how learning was measured could influence the magnitude of the results observed; however, for this tool, we accepted the authors’ assertions at face value that these were relevant proxies for learning and achievement. We do not claim that this literature represents all the published research-based best practices for implementing active learning in the classroom. Instead, these are the baseline recommendations, and, as more studies are published we will continue to modify PORTAAL to reflect the latest evidence-based teaching practices for active learning.

The majority of the articles reporting on active learning in the DBER literature involve whole-class transformations in which multiple features differ between the control and experimental classes. Although important, these studies were not useful for identifying specific features correlated with increases in student outcomes, because no one feature was tested in isolation. For example, an intervention by Freeman et al. (2011) increased the active learning in class by 1) adding volunteer and cold-call discussions, 2) providing reading quizzes before class, and 3) providing weekly practice exams. With all these changes, it was impossible to determine exactly which component led to the observed increase in academic achievement. Thus, articles of this type were not used in the development of PORTAAL. Instead, we focused on articles like the one by Smith et al. (2011) that explicitly test one element of the implementation of active learning (in this case, the role of peer discussions) to determine the impact of that single element on student outcomes.

Our survey of the research literature found that best practices for implementing active learning clustered along four dimensions: 1) practice, 2) logic development, 3) accountability, and 4) apprehension reduction. The first two dimensions deal with creating opportunities for practice and the skills this practice reinforces. The second pair addresses how to encourage all students to participate in that practice.

**Dimension 1: Practice**. The first dimension in the PORTAAL rubric is a measure of the amount and quality of practice during class (Table 1). There are many articles validating the importance of opportunities to practice: student learning is positively correlated with the number of in-class clicker questions asked (Preszler et al., 2007); students who created their own explanations performed better on related exam questions than students who read expert explanations (Wood et al., 1994; Willoughby et al., 2000); and repeated practice testing is correlated with both increased learning (Dunlosky et al., 2013) and student metacognition (Thomas and McDaniel, 2007). Thus, best practice is to provide opportunities in class for students to practice (PORTAAL element practice 1, P1).

In addition to the amount of practice, the quality and distribution of the practice is also important (Ericsson et al., 1993). For practice to increase achievement, the practice must be similar to the tasks students are expected to perform (transfer-appropriate principle; Morris et al., 1977; Ericsson et al., 1993; Thomas and McDaniel, 2007; Jensen et al., 2014). One method of measuring this alignment of practice and assessment is to determine how similar exam questions are to in-class questions. Jensen et al. (2014) provide a dramatic demonstration of the importance of this alignment: when students engage in higher-order skills in class but exams only test low-level skills, students fail to acquire higher-level skills. These results reinforce the concept that the test greatly influences what the students study. Additional studies have supported this finding that students learn what they are tested on (Morgan et al., 2007; Wormald et al., 2009). In PORTAAL, we use the cognitive domain of Bloom’s taxonomy to determine the alignment between in-class practice and exams (P2; cf. Crowe et al., 2008).
Table 1. Elements in the dimension of practice and the evidence supporting them*

<table>
<thead>
<tr>
<th>Elements</th>
<th>How element is observed in the classroom</th>
<th>Increases achievement</th>
<th>Improves conversations</th>
<th>Improves other measures</th>
<th>Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1. Frequent practice</td>
<td>Minutes any student has the possibility of talking through content in class</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>Wood et al., 1994; Willoughby et al., 2000; Preszler et al., 2007; Thomas and McDaniel, 2007; Dunlosky et al., 2013 (review)</td>
</tr>
<tr>
<td>P2. Alignment of practice and assessment</td>
<td>In-class practice questions at same cognitive skills level as course assessments (requires access to exams)</td>
<td>✓</td>
<td></td>
<td>McDaniel et al., 1978; Morris et al., 1977; Thomas and McDaniel, 2007; Ericsson et al., 1993; Jensen et al., 2014; Wormald et al., 2009; Morgan et al., 2007</td>
<td></td>
</tr>
<tr>
<td>P3. Distributed practice</td>
<td>Percent of activities in which instructor reminds students to use prior knowledge</td>
<td>✓</td>
<td></td>
<td>deWinstanley et al., 2002; Dunlosky et al., 2013</td>
<td></td>
</tr>
<tr>
<td>P4. Immediate feedback</td>
<td>Percent of activities in which instructor hears student logic and has an opportunity to respond</td>
<td>✓</td>
<td></td>
<td>Renkl, 2002; Epstein et al., 2002; Ericsson et al., 1993; Trowbridge and Carson, 1932</td>
<td></td>
</tr>
</tbody>
</table>

*Measures are positively correlated with dimension unless otherwise stated. All these measures were on adult learners, although they were not all in large-lecture contexts.

Finally, the temporal spacing of practice is important. Cognitive psychologists have shown that practice that is spaced out is more effective than massed practice on a topic (distributed practice; deWinstanley and Bjork, 2002; Dunlosky et al., 2013). These findings imply instructors should consider designing activities that revisit a past topic or ask students to relate the current topic to a past topic. Classroom observers could detect this distributed practice when instructors explicitly cue students to use their prior knowledge (P3).

Immediate feedback and correction also improves student performance (Trowbridge and Carson, 1932; Ericsson et al., 1993; Epstein et al., 2002; Renkl, 2002). In large lectures, this is primarily accomplished when students provide explanations for their answers to the instructor in front of the class as a whole (P4).

Dimension 2: Logic Development. The second dimension in the PORTAAL rubric is a measure of the development of higher-order thinking skills (Table 2). Few articles validate the effect of this dimension on changes in achievement, because most exams test low-level questions (Momsen et al., 2010). However, there is extensive evidence documenting how this dimension increases the quality of student conversations or other relevant measurements, such as changes in goal orientation (i.e., students focusing more on understanding the material than on getting the correct answer).

To provide students with opportunities to practice their logic development, it is necessary for instructors to formulate questions that require a higher level of thinking (PORTAAL element logic development 1, L1; Morris et al., 1977; Ericsson et al., 1993; Jensen et al., 2014). One method for writing questions that require logic and critical thinking is to specifically write questions at higher Bloom levels (cf. Crowe et al., 2008).

One of the simplest ways to improve student conversations and increase student focus on sense-making and logic when introducing activities is for instructors to remind students to provide the rationale for their answers (L2; Turpen and Finkelstein, 2010; Knight et al., 2013). Additional evidence for the importance of encouraging students to explain their logic comes from Willoughby et al. (2000) and Wood et al. (1994), who demonstrated that students who generate explanations perform better on exams than those who did not.

When students begin to work on an instructor-posed question (such as a clicker question), the inclusion of several elements can increase student outcomes related to logic development. First, providing an explicitly delineated time at the beginning of the discussion period when students have an opportunity to think through their answers on their own (L3) increases the likelihood that a student will choose to join in a subsequent small- or large-group discussion and thus improve the discussion (Nielsen et al., 2012). Students also self-report that having time to think independently before discussions allows time to think through the question and come up with their own ideas (Nicol and Boyle, 2003). This individual time can be set up as a short minute-writing exercise or as a clicker question answered individually.

Second, a deeper understanding of the material is often gained when students share and explain their answers to other students (i.e., work in small groups; L4; Wood et al., 1995, 1999; Renkl, 2002; Sampson and Clark, 2009; Menekse et al., 2013). This group work increases performance on isomorphic clicker questions and posttests compared with performance of students who only listened to an instructor explanation (Schworm and Renkl, 2006; Smith et al., 2011). Group work is particularly important for tasks that require students to transfer knowledge from one context to another (Gadgil and Nokes-Malach, 2012). Small-group work also increased the frequency of students spontaneously providing
Table 2. Elements in the dimension of logic development and the evidence supporting them

<table>
<thead>
<tr>
<th>Dimension 2: Logic Development</th>
<th>How element is observed in classroom</th>
<th>Increases achievement</th>
<th>Improves conversations</th>
<th>Improves other measures</th>
<th>Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1. Opportunities to practice higher-order skills in class</td>
<td>Percent of activities that require students to use higher-order cognitive skills</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Jensen et al., 2014; Ericsson et al., 1993; Morris et al., 1977</td>
</tr>
<tr>
<td>L2. Prompt student to explain/defend their answers</td>
<td>Percent of activities in which students are reminded to use logic</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Smith et al., 2011; Sampson and Clark, 2009; Len, 2006–2007; Gadgil and Nokes-Malach, 2012; Wood et al., 1994; Wood et al., 1999; Menekse et al., 2013; Schworm and Renkl, 2006; Len, 2006–2007; Hoekstra and Mollborn, 2011; Okada and Simon, 1997</td>
</tr>
<tr>
<td>L3. Allow students time to think before they discuss answers</td>
<td>Percent of activities in which students are explicitly given time to think alone before having to talk in groups or in front of class</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Perez et al., 2010; Brooks and Konesky, 2011; Nielsen et al., 2012; Kulhavy, 1977</td>
</tr>
<tr>
<td>L4. Students explain their answers to their peers</td>
<td>Percent of activities in which students work in small groups during student engagement</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Smith et al., 2011; Sampson and Clark, 2009; Len, 2006–2007; Gadgil and Nokes-Malach, 2012; Wood et al., 1994; Wood et al., 1999; Menekse et al., 2013; Schworm and Renkl, 2006; Len, 2006–2007; Hoekstra and Mollborn, 2011; Okada and Simon, 1997</td>
</tr>
<tr>
<td>L5. Students solve problems without hints</td>
<td>Percent of activities in which answer is not hinted at between iterations of student engagement</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Smith et al., 2011; Butler et al., 2013; Nicol and Boyle, 2003 (S); Nielsen et al., 2012 (S); Turpen and Finkelstein, 2010 (S)</td>
</tr>
<tr>
<td>L6. Students hear students describing their logic</td>
<td>Percent of activities in which students share their logic in front of the whole class</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Smith et al., 2011; Butler et al., 2013; Nicol and Boyle, 2003 (S); Nielsen et al., 2012 (S)</td>
</tr>
<tr>
<td>L7. Logic behind correct answer explained</td>
<td>Percent of activities in which correct answer is explained</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Smith et al., 2011; Butler et al., 2013; Nicol and Boyle, 2003 (S); Nielsen et al., 2012 (S)</td>
</tr>
<tr>
<td>L8. Logic behind why incorrect or partially incorrect answers are explained</td>
<td>Percent of activities in which alternative answers are discussed during debrief</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Nielsen et al., 2012 (S); Turpen and Finkelstein, 2010 (S)</td>
</tr>
</tbody>
</table>

*Measures are positively correlated with dimension unless otherwise stated. Citations with an (S) are student self-reported measures. All these measure were on adult learners (unless denoted with an asterisk), although they were not all in large-lecture contexts.

explanations for their answers (Okada and Simon, 1997). In addition, small-group discussions 1) improve student attitudes, particularly the attitudes of low-performing students, toward the discussion topic (Len, 2006–2007); 2) improve social cohesion and, through that, feelings of accountability (Hoekstra and Mollborn, 2011); and 3) develop students’ argumentation skills (Kuhn et al., 1997; Eichinger et al., 1991).

Another key element involves the instructor hinting at or revealing correct answers. Correct answers should not be hinted at between iterations of discussion (L5). For example, showing the histogram of clicker responses can cause more students to subsequently choose the most common answer whether or not it is right (Perez et al., 2010; Brooks and Konesky, 2011; Nielsen et al., 2012). Seeing the most common answer also leads to a reduction in the quality of peer discussions (Nielsen et al., 2012). In addition, making the correct answer easier to access (through hints or showing a histogram of student responses) reduces the effort students put into determining the correct answer and therefore reduces learning outcomes (Kulhavy, 1977).

It is also important that the whole class hear a fellow student offer an explanation for the answer selected (L6). This provides all students with immediate feedback on their logic and critical thinking, not just the right answer (Turpen and Finkelstein, 2010). Although not explicitly tested at the college level, an instructor asking students to provide their logic in whole-class discussions has shown the additional benefit of increasing the amount of explaining students do in small-group discussions in K–12 classes (Webb et al., 2008, 2009).

Finally, it is important for the instructor to explain why the correct answer is correct (L7; Smith et al., 2011; Nielsen et al., 2012). This provides a model for students of the type of logical response expected. Some students consider this
Table 3. Elements in the dimension of accountability and the evidence supporting them

<table>
<thead>
<tr>
<th>Elements</th>
<th>How element is observed in the classroom</th>
<th>Increases achievement</th>
<th>Improves conversations</th>
<th>Improves other measures</th>
<th>Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1. Activities worth course points</td>
<td>Percent activities worth course points (may require a syllabus or other student data source)</td>
<td>✓</td>
<td>✓</td>
<td>✓ participation 1–6</td>
<td>¹Freeman et al., 2007; ²Len, 2006–2007; ³Willoughby and Gustafson, 2009; ⁴Perez et al., 2010; ⁵James and Willoughby, 2011; ⁶James, 2006</td>
</tr>
<tr>
<td>A2. Activities involve small-group work, so more students have opportunity to participate</td>
<td>Percent activities in which students work in small groups</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Hoekstra and Mollborn, 2011; Chidambaram and Tung, 2005; Aggarwal and O’Brien, 2008</td>
</tr>
<tr>
<td>A3. Avoid volunteer bias by using cold call or random call</td>
<td>Percent activities in which cold or random call used</td>
<td>✓</td>
<td></td>
<td></td>
<td>Dallimore et al., 2013; Eddy et al., 2014</td>
</tr>
</tbody>
</table>

*Measures are positively correlated with dimension unless otherwise stated. All these measures were on adult learners, although they were not all in large-lecture contexts.

Dimension 3: Accountability. Motivating students to participate is another critical component for the success of any active-learning classroom. This motivation could come through the relevancy of the activity (Bybee et al., 2006), but this relevancy could be challenging for an observer to know, so it was excluded from our tool. Instead, we focused on teacher-provided incentives that create attention around the activity (Table 3; deWinstanley and Bjork, 2002). One such incentive is to make activities worth course points (PORTA-AL element accountability 1, A1). Awarding course points has been shown to increase student participation (especially the participation of low-performing students; Perez et al., 2010) and increase overall class attendance and performance (Freeman et al., 2007). There are two primary strategies used by instructors for assigning course points for in-class activities: participation and correct answer. Having students earn points only for correct answers has been shown to increase student performances on these questions and in the course (Freeman et al., 2007), but it has also been shown to reduce the quality of the small-group discussions, as 1) these interactions were dominated by the student in the group with the greatest knowledge; 2) there were fewer exchanges between students, and 3) fewer students participated in the discussions overall (James, 2006; James and Willoughby, 2011). In an experiment explicitly designed to test how grading scheme impacts learning, Willoughby and Gustafson (2009) found that although students answered more clicker questions correctly when points were earned based on correctness, this grading scheme did not lead to an increase in achievement on a posttest relative to the participation-point condition. This result was supported by the Freeman et al. (2007) study as well. These results, paired with the other studies, demonstrate that both conditions lead to equal learning but that the participation-point condition leads to more complex student discussions. Instructors should thus choose the outcome most important to them and use the associated grading scheme.

A second way to motivate student participation is by creating situations wherein students will have to explain their responses either in small groups (A2) or to the whole class (A3). However, most instructors elicit responses from students by calling on volunteers in the classroom setting (Eddy et al., 2014). Volunteer responses are generally dominated by one or a few students, and the majority of the class know they will not be called on and thus do not need to prepare to answer (Fritschner, 2000). Volunteer-based discussions also demonstrate a gender bias in who participates, with males speaking significantly more than females (Eddy et al., 2014). For these reasons, the research literature recommends instructors find alternative methods for encouraging students to talk in class.

Alternatives to volunteer responses include: small-group work (A2) and random or cold-calling (A3). Although participating in small-group work may still seem voluntary, psychology studies in classrooms have demonstrated that the smaller the group a student is in (e.g., pairs vs. the entire class) the more likely he or she is to participate in group work and the higher the quality of individual answers (Chidambaram and Tung, 2005; Aggarwal and O’Brien, 2008). The idea behind this pattern is that there is a dilution of responsibility and reward in a large class (students are fairly anonymous and will not get personally called out for not participating, and, even if they do participate, the chance they will be rewarded by being called on is low), which decreases motivation to participate (Kidwell and Bennett, 1993). In small groups, the situation is different: student effort (or lack thereof) is noticed by group mates. In addition, social cohesion, attachment to group mates, is also more likely to form between members of small groups than between members of a large lecture class, and this cohesion increases a student’s sense of accountability (Hoekstra and Mollborn, 2011).
Cold-calling involves calling on students by name to answer a question. Random call is a modified version of cold-calling in which instructors use a randomized class list to call on students. Although these methods may seem intimidating and punitive to students, researchers have shown that cold-calling, when done frequently, actually increases student self-reported comfort speaking in front of the class and is correlated with students becoming more willing to volunteer in class (Dallimore et al., 2013). In addition, random call both eliminates gender bias in participation (Eddy et al., 2014) and guarantees all students have an equal chance of being called on.

**Dimension 4: Reducing Student Apprehension.** The final dimension is also related to increasing student motivation to participate in class. Instead of raising the incentive to participate, instructors can frame their activities in ways that reduce a student’s fear of participation. This can be done in a number of ways, but many of these would be hard to reliably document, so we focus instead on three strategies that are explicit and observable and have shown positive changes correlated with student learning and/or participation (Table 4).

One of the most common ways instructors motivate students to participate is through confirmation. Confirmation behaviors are those that communicate to students that they are valued and important (Ellis, 2000). These instructor behaviors have been correlated with both affective learning (how much students like a subject or domain) and the more typical cognitive learning (Ellis, 2000, 2004; Goodboy and Myers, 2008). Behaviors students generally interpret as affiliating include but are not limited to: 1) the instructor not focusing on a small group of students and ignoring others, which random call accomplishes (PORTAAL element apprehension reduction, R1); or praising the efforts of the whole class rather than an individual student (R2); and 2) communicating support for students by indicating that student comments are appreciated (praising a student’s contribution, R3) and not belittling a student’s contribution (R4). These behaviors can also influence students’ willingness to participate in class (Fritschner, 2000; Goodboy and Myers, 2008).

In addition to using confirmation behaviors, instructors can increase participation by framing 1) student mistakes as natural and useful and/or 2) student performance as a product of their effort rather than their intelligence. The first type of framing, called error framing (R5), increases student performance by lowering anxiety about making mistakes (Bell and Kozlowski, 2008). The framing itself is simple. In the Bell and Kozlowski (2008) study, students in the treatment group were simply told: “errors are a positive part of the training process” and “you can learn from your mistakes and develop a better understanding of the [topic/activity].” Emphasizing errors as natural, useful, and not something to be afraid of can also encourage students to take risks and engage in classroom discussions.

---

<table>
<thead>
<tr>
<th>Elements</th>
<th>How element is observed in the classroom</th>
<th>Increases achievement</th>
<th>Improves conversations</th>
<th>Improves other measures</th>
<th>Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1. Give students practice participating by enforcing participation through cold/random call</td>
<td>Percent activities with random or cold-calling used during student engagement or debrief</td>
<td>√</td>
<td>1–3</td>
<td>√</td>
<td>Ellis, 2004 (S); Dallimore et al., 2010</td>
</tr>
<tr>
<td>R2. Student confirmation: provide praise to whole class for their work</td>
<td>Percent debriefs and engagements in which class received explicit positive feedback and/or encouragement</td>
<td>1–3</td>
<td>3, 4</td>
<td>3, 4</td>
<td>1Ellis, 2004 (S); 2Ellis, 2000 (S); 3Goodboy and Myers, 2008 (S); 4Fritschner, 2000</td>
</tr>
<tr>
<td>R3. Student confirmation: provide praise/encouragement to individual students</td>
<td>Percent student responses with explicit positive feedback and/or encouragement</td>
<td>1–3</td>
<td>3, 4</td>
<td>3, 4</td>
<td>1Ellis, 2004 (S); 2Ellis, 2000 (S); 3Goodboy and Myers, 2008 (S); 4Fritschner, 2000</td>
</tr>
<tr>
<td>R4. Student confirmation: do not belittle/insult student responses</td>
<td>Percent student responses that do not receive negative feedback</td>
<td>√</td>
<td>1</td>
<td>2</td>
<td>1Ellis, 2004 (S); 2Fritschner, 2000</td>
</tr>
<tr>
<td>R5. Error framing: emphasize errors natural/instructional</td>
<td>Percent activities in which instructor reminds students that errors are nothing to be afraid of during introduction or student engagement periods</td>
<td>√</td>
<td></td>
<td></td>
<td>Bell and Kozlowski, 2008</td>
</tr>
<tr>
<td>R6. Emphasize hard work over ability</td>
<td>Percent activities in which instructor explicitly praises student effort or improvement</td>
<td>√</td>
<td></td>
<td></td>
<td>Aronson et al., 2002; Good et al., 2012</td>
</tr>
</tbody>
</table>

*Citations with an (S) are student self-reported measures. Measures are positively correlated with dimension unless otherwise stated. All these measure were on adult learners, although they were not all in large-lecture contexts.*
The second type of framing, the growth mind-set, is intended to change a student’s course-related goals. There is considerable evidence in the K–12 literature that students who hold a performance goal (to get a high grade in the course) will sacrifice opportunities for learning if those opportunities threaten their ability to appear “smart” (Elliot and Dweck, 1988). This risk avoidance manifests as not participating in small-group work or not being willing to answer instructor questions if the questions are challenging. One way to influence student goals for the class is through praise based on effort rather than ability. Praise based on ability (“being smart”) encourages students to adopt a performance goal rather than a mastery goal (a desire to get better at the task) and to shy away from tasks on which they might not perform well. Praise that focuses on effort encourages a mastery goal and can increase student willingness to participate in challenging tasks (R6; Mueller and Dweck, 1998). At the college level, an intervention that had students reflect on a situation in their own life and Dweck, 1998). This risk avoidance manifests as not participating in small-group work or not being willing to answer instructor questions if the questions are challenging. One way to influence student goals for the class is through praise based on effort rather than ability. Praise based on ability (“being smart”) encourages students to adopt a performance goal rather than a mastery goal (a desire to get better at the task) and to shy away from tasks on which they might not perform well. Praise that focuses on effort encourages a mastery goal and can increase student willingness to participate in challenging tasks (R6; Mueller and Dweck, 1998). At the college level, an intervention that had students reflect on a situation in their own life and to consider a future situation in which they felt successful because of their own actions (Aronson et al., 2002). Furthermore, for some college students, the perception that their instructor holds mastery goals for the class leads to improved performance relative to students who believe that their instructor holds performance goals (Good et al., 2002). 

In summary, we identified 21 elements that capture four dimensions of best practices for active learning (practice: P1–4; logic development: L1–8; accountability: A1–3; and apprehension reduction: R1–6). 

Goal 3: Validity, Reliability, and Ease of Use of PORTAAL

Content Validity. The validity of PORTAAL is based in the research literature. Each of the 21 elements is supported by at least one published peer-reviewed article that demonstrates its impact on a relevant student outcome (see Goals 1 and 2: Identifying Research-Based Best Practices for Large Classrooms (Literature Review). As more research is done on the finer points of active-learning methods in the classroom, we will continue to modify PORTAAL to keep it up to date.

Face Validity. After the elements were identified from the research literature, we discussed how the elements might be observed in the classroom. The observations we developed are not necessarily perfect measures but are nevertheless indicators for the potential presence of each element. Furthermore, the defined observations met our goal that observers without deep pedagogical content knowledge can use it readily and reliably.

Though there are 21 elements in PORTAAL, three of the elements could be consolidated into other elements for an online face-validity survey (see the Supplemental Material). We presented these 18 observations to seven BER researchers who have published in the field within the past 2 yr. We asked these BER researchers whether they agreed or disagreed that the stated observation could describe the presence of a particular element in the classroom. If the reviewer disagreed, we asked him or her to indicate why. There was 100% agreement for 11 of the 18 observations. Seven elements had 86% agreement, and one element had 71% agreement (see Validity (DBER researcher agreement) column in Supplemental Table 2: BER Review Comments for more details).

Reliability. PORTAAL observations are completed on an observation log (see https://sites.google.com/site/uwbioedresgroup/research/portaal-resources) that asks observers to record the timing, frequency, and presence/absence of events as they occur in an activity. For example, observers record the start and end time of a small-group discussion, they count the number of students who talk during an activity debrief, and they record whether or not students are explicitly reminded to explain their answers to one another. An activity was operationally defined as any time students engage with a novel problem or question. An activity can be challenging to delineate if there are a series of questions in a row. We enforced the rule that if the next question is beyond the scope of the initial question, then it is a new activity. Ultimately, activity characteristics are pooled to create an overall description of the class session. Our aim with PORTAAL was that the recording of these discrete and observable elements would make this tool reliable and easy to learn. In this section, we test whether we accomplished this aim.

PORTAAL observations are done in pairs. Each observer independently watches a class session and records his or her observation on the log, and then observers come to consensus on what they observed. In our study, we used two individuals who had no prior teaching experience (a recent graduate from the undergraduate biology program and a master’s student in engineering) to observe 15 different class sessions. To determine interobserver reliability, we analyzed their original independent observations before they came to consensus. We used the intraclass correlation coefficient (ICC) to assess reliability data, as our data were interval data (percentages). We used a two-way agreement, single-measures ICC (McGraw and Wong 1996). The resulting ICCs were all > 0.8, and thus all were in the “excellent” range (Supplemental Table 2; Cicchetti 1994). This indicated a high degree of agreement across both our coders and supports the reliability of the PORTAAL measures. 

Next, to determine how easy it was for different types of observers to learn PORTAAL, we trained and assessed the observations of four additional sets of observers. One pair of observers were current undergraduates in fields other than biology with no teaching experience (individuals without deep pedagogical content knowledge or disciplinary content knowledge). Another set was a graduate student in biology and a postdoc who had some teaching experience. The final two sets of observers were instructors with extensive teaching experience.

To train these four pairs of observers, we developed a training manual and a set of three short practice videos (available at https://sites.google.com/site/uwbioedresgroup/research/portaal-resources). Training involves a moderate time commitment that can occur over multiple days: 1 h to read the manual, 1 h to practice Blooming questions and come to consensus with a partner, and 2–3 h of practice scoring sample videos and reaching consensus.

We compared the scores of the four training teams with the scores of our experienced pair of observers for the first training video to get a sense of how quickly new observers could...
learn PORTAAL. Although statistics could not be run on such small sample sizes, we found that, on their first attempt to score a video, the four pairs of novice observers exactly matched the expert ratings ≥ 90% of the time for 19 of the 21 dimensions of PORTAAL. There were only two dimensions with less than an 85% match with the experts. These two dimensions were Bloom level of the question (matched 80% of the activities) and whether or not an instructor provided an explanation for the correct answer (70% of the activities). These data validate that the discrete and observable elements in PORTAAL make this tool reliable and easy to learn.

PART 2: ASSESSING CLASSROOM PRACTICES WITH PORTAAL

Is There Variation in the Use of Active-Learning Best Practices in Large Lectures?

The use of active learning in STEM disciplines usually increases student achievement (Freeman et al., 2014), but the extent of instructor implementation of those active-learning strategies in the classroom varies. In this section, we test whether PORTAAL can document this variation.

Our Sample

We used PORTAAL to document the teaching practices of 25 instructors teaching in the three-quarter introductory biology series at a large public R1 university. The first course in the series focuses on evolution and ecology; the second on molecular, cellular and developmental biology; and the third on plant and animal physiology. The majority (n = 21) of the instructors in this study only taught half of the 10-wk term, while four taught the whole term. Classes ranged in size from 159 to more than 900 students. The instructors in this sample all used clickers or other forms of student engagement.

One of the strengths of this study is its retrospective design. At this university, all courses are routinely recorded to allow students in the classes to review the class sessions. We used this archived footage, rather than live observations, to preclude instructors from changing their teaching methods. The archived footage also allows observers to pause the video while they record observations. We do not believe it will ever be practical to implement PORTAAL in real time, so recordings are critical for effective use of this tool. Our classroom recordings were done from the back of the room, so we had a view of the instructors as they moved around the front of the room, the screen on which they projected their PowerPoints, and a large section of the students in the class. The sound in our recordings came from the microphone the instructor used, so we generally could hear what the instructor could hear.

Kane and Cantrell (2013) found that two trained individuals observing the same 45-min session of a teacher’s class captured instructor teaching methods as well as having four independent observations of four different class sessions, indicating that fewer classes can be observed if two observers are used. Based on this finding, in our study, to be conservative and to increase the number of student–teacher interactions sampled, we decided to observe three randomly selected class sessions from across the term for each instructor. These 75 videos were first independently scored by two individuals (a recent biology graduate and an engineering master’s student) using the PORTAAL observation log; this was followed by the observers reaching consensus on all observations.

In addition to the 25 instructors, we identified two instructors whose implementations of active learning have been documented to increase student achievement on instructor-written classroom exams. Bloom levels of exam questions were assessed, and exams were determined to be equivalent, if not a little more challenging, in the reformed course across the terms included in these studies (for details, see Freeman et al., 2011; Eddy et al., 2014). These two instructors changed aspects both inside and outside their classrooms, and, at this time, we cannot fully resolve what proportion of the student learning gains are due to the instructor’s classroom practices. However, we were interested to see whether these instructors were more frequently employing PORTAAL elements in their classrooms. We will call these instructors our “reference” instructors for the remainder of this paper. The data from these instructors are included on Figures 1–5 as possible targets for instructors who desire to achieve gains in student achievement. The same methods as those described above were used to score three videos of each of these two reference instructors.

The blue reference instructor teaches the first course in an introductory biology sequence for mixed majors. The course covers general introductions to the nature of science, cell biology, genetics, evolution and ecology, and animal physiology, and averages more than 300 students a term. The blue instructor decreased failure rates (instances of a “C−” or lower) on exams by 41% (Eddy and Hogan, 2014) when the instructor changed from traditional lecturing to a more active-learning environment that incorporated greater use of group work in class and weekly reading quizzes.

The red reference instructor teaches the first course in an introductory biology sequence for biology majors. This course covers topics including evolution and ecology. The course ranges from 500 to 1000 students a term. This instructor has changed the classroom environment over time to incorporate more group work, more opportunities for in-class practice on higher-order problems, and less instructor explanation. Outside-of-class elements that changed included the addition of reading quizzes and practice exams. With these changes, the failure rate decreased dramatically (a 65% reduction; Freeman et al., 2011). Again, we cannot specifically identify the relative contribution of the change in classroom practices to the change in failure rate, but we can say it was one of three contributing elements.

Calculating PORTAAL Scores

The 21 elements in PORTAAL do not sum to a single score, as we do not know the relative importance of each element for student outcomes. Thus, instead of a score, instructors receive the average frequency or duration of each element across their three class sessions (the conversion chart can be found at https://sites.google.com/site/uwbioedresgroup/research/portal-resources). We hypothesize that the more frequently instructors employ the elements documented in PORTAAL, the more students will learn. We test this hypothesis visually.
in a very preliminary way by looking at where the two reference instructors fall on each dimension. We do not expect these two instructors to practice all the elements, but we do predict on average that they will incorporate the elements more frequently.

To identify the variation in implementation of each element across the 25 instructors, we calculated and plotted the median and quartiles for each element. Thus, instructors are able to compare their PORTAAL element scores with this data set and select elements they wish to target for change.

**Results**

Overall, we found large variation in how instructors implement active learning in introductory biology classrooms. In addition, we see that both reference faculty members had values in the top (fourth) quartile for 52% of the PORTAAL elements, suggesting that more frequent use of PORTAAL elements increases student learning.

**Dimension 1: Practice.** More than half the instructors allowed students <6 min per 50-min class session to engage in practice, whereas the two reference instructors allowed 17 and 31 min (P1; Figure 2A). Evidence for the use of distributed practice was low, with a median of 4.2% of activities explicitly cueing students to use prior knowledge, whereas the two reference instructors had ~13% of activities focused on this (P3; Figure 2B). The element of immediacy of feedback on student ideas occurred much more frequently than any of the other dimensions of practice: the median number of activities in which instructors heard student explanations was 60%, and instructors in the third and fourth quartiles overlapped with the reference instructors (84.7% and 76.5%; P4; Figure 2C). We could not assess the alignment of practice (P2), as we did not have exams for all these instructors.

**Dimension 2: Logic Development.** Logic development had the greatest variation of the four dimensions, with some elements performed routinely and others not at all. Instructors routinely followed best practices and did not show a clicker histogram of results or give students hints between iterations of student engagement (median 96.1% of activities; L5; Figure 3E), and instructors routinely explained why the correct answer was right during debriefs (71.8% of activities; L7; Figure 3F). Values for both these elements were similar in frequency to those of our reference instructors. Other elements of logic development were less evident: a median of 15% of activities involved higher-order cognitive skills (L1), whereas our reference instructors used higher-order questions in 34.7 and 64.2% of the activities (Figure 3A); 32.2% of activities involved students talking through their answers in small groups (L4), which was similar to the blue instructor but half that of the red instructor (Figure 3D); 8.3% of activities initiated student engagement with an initial think-alone period for students to formulate their answers (L3), which was less than either reference instructor (Figure 3C); and 10.3% of activities involved debriefs with explanations for why the wrong answers were wrong (L8), which was similar to the reference instructors (Figure 3G). Finally, in our sample, less than 25% of instructors explicitly reminded students to provide the logic for their answers when they introduced the activity (L2; Figure 3B). This was true for the reference instructors as well.

**Dimension 3: Accountability.** On average, the median percent of activities with at least some form of accountability (points, random call, or small-group work) across our sample was high: 65.5%. Most of that accountability comes from assigning course points to engagement (22.7% of activities; A1; Figure 4A) and small-group work (A2; Figure 4D). Fewer than 25% of instructors use cold or random call in their courses (median 0% of activities; A3), including the blue instructor (Figure 4B). The red instructor uses random call during 56% of activities.

**Dimension 4: Apprehension Reduction.** In general, instructors (including reference instructors) did not direct much effort toward reducing student’s apprehension around class participation in the observed videos: the median percent of
hard work), so we cannot provide specific numbers for these elements. Anecdotally, neither observer believed they observed any explicit incidences of either.

**Implications of PORTAAL Results**

These PORTAAL results indicate there is considerable variation in instructor implementation of evidence-based best teaching practices in the classroom in introductory

---

**Figure 3.** Dimension 2: Logic development—variation in implementation of elements. Histograms demonstrating the variation in instructor classroom practice for each element of the dimension of logic development. The black dotted line is the median for the 25 instructors; the red line is the practice of the instructor who reduced student failure rate by 65%; and the blue line is the instructor who reduced failure rate by 41%. Each quartile represents where the observations from 25% of the instructors fall. Quartiles can appear to be missing if they overlap with one another.
ability by using cold call or random call more frequently; and 4) reduce participation apprehension by reminding students that mistakes are part of learning and not something to be feared. Obviously, an instructor could not address all of these at once but could choose which to pursue based on his or her own interests and PORTAAL results.

In addition, PORTAAL can be used to identify how instructor teaching practices differ from one another. By comparing our two reference instructors with the 25 instructors and with one another, we see that the reference instructors employ more PORTAAL elements more frequently than the median of the instructors in our sample. In addition, the red instructor, who had the greatest decrease in failure rate, used several PORTAAL elements (including P1, L1, L3, and L4) more frequently than the blue instructor. This again suggests that more frequent use of PORTAAL elements increases student learning.

In our sample, we see that instructors, on average, frequently practice many of the PORTAAL elements: they do not regularly provide hints during the engagement part of the activity, do frequently explain why the right answer is correct, do make students accountable for participating, and do not discourage student participation with negative feedback. Our results also lead us to make specific recommendations for improvement: instructors in this sample could 1) increase opportunities for practice by incorporating more distributed practice and more higher-order problems in class; 2) improve the development of student logical-thinking skills by reminding students to explain their answers, providing students explicit opportunities to think before they talk, and using random call to spread participation across the class; 3) increase participation through account-

Figure 4. Dimension 3: Accountability—variation in implementation of elements. Histograms demonstrating the variation in instructor classroom practice for each element of the dimension of accountability. The black dotted line is the median for the 25 instructors; the red line is the practice of the instructor who reduced student failure rate by 65%; and the blue line is the instructor who reduced failure rate by 41%. Each quartile represents where the observations from 25% of the instructors fall. Quartiles can appear to be missing if they overlap with one another.

Figure 5. Dimension 4: Apprehension reduction—variation in implementation of elements. Histograms demonstrating the variation in instructor classroom practice for each element of the dimension of apprehension reduction. The black dotted line is the median for the 25 instructors; the red line is the practice of the instructor who reduced student failure rate by 65%; and the blue line is the instructor who reduced failure rate by 41%. Each quartile represents where the observations from 25% of the instructors fall. Quartiles can appear to be missing if they overlap with one another.
Limitations of PORTAAL
PORTAAL, like any tool, was designed with a particular purpose and scope: to reliably evaluate the alignment between instructor implementations of active learning and research-supported best practices in the classroom. Thus, the focus of the tool is on classroom practice, but the effectiveness of an active-learning classroom also depends on elements not captured by the tool, including characteristics of the students, characteristics of exams, course topics, and student activities outside class (Figure 1). For example, it is likely that the usefulness of what goes on in class is determined by how well students are prepared for class and how much practice they have outside of class. These out-of-class elements could in turn be influenced by student characteristics such as their prior academic preparation and their motivation (Figure 1). PORTAAL does not capture outside-class assignments or student-level characteristics. This is an important caveat, as most of the studies in this review were conducted at selective R1 universities that have a very specific population of students. In addition, we explicitly chose not to focus on how the instructor scaffolds course content. This is a critical aspect of instruction, but it is difficult to assess without deep pedagogical content knowledge. Finally, PORTAAL focuses on observable and explicit behaviors in the classroom. Some elements may not be perfectly captured by these types of measures. For example, teacher confirmation behaviors are only measured by explicit instances of praise, but there are many other ways that teachers can confirm students, such as body language or tone of voice. For a more nuanced picture of instructor confirmation, instructors could use the survey developed by Ellis (2000). In addition, our measure of distributed practice can only measure instances in which the instructor explicitly cues students to use prior knowledge, which will likely underestimate instances of distributed practice. These limitations are necessary for reliability in the instrument but may lead to underestimating the frequency of the element. The final limitation of PORTAAL is that it assumes the observers can record all the important interactions that go on in the classroom. This limits the classroom types this tool can effectively evaluate. PORTAAL is designed for large-enrollment courses in which it would be difficult for an instructor to interact individually with the majority of the students in a class period. PORTAAL would not work well in a small seminar-style discussion with frequent student-to-student, whole-class discussion or a lab course in which students are at different stages of a project. In addition, it is not feasible to use PORTAAL reliably in real time. We recommend videotaping the class with a focus on the instructor and a view of some of the students. Despite these limitations, we see this rubric as useful for the majority of STEM classes. Following the suggestions outlined in this tool does not guarantee greater student learning, but the tool is a solid, research-supported first step.

PART 3: CONCLUSION

How Can PORTAAL Increase Implementation of Evidence-Based Active-Learning Activities in the Classroom?

From our analysis of 25 instructors, it is evident there is extensive variation in implementing research-based best practices for active learning in the classroom. This result is not exclusive to biology, as it has also been seen in engineering (Borrego et al., 2013), physics (Henderson and Dancy, 2009), and extensively at the K–12 level (O’Donnell, 2008). One of the major causes of such variation may be the nebulous nature of active learning. One of the most popular definitions of active learning comes from Hake (1998): “those [practices] designed at least in part to promote conceptual understanding through interactive engagement of students in heads-on (always) and hands-on (usually) activities which yield immediate feedback through discussion with peers and/or instructors.” Although a good summary, this statement does not provide practical guidance on the proper implementation of active learning.

Unfortunately, the education literature does not always identify the key elements of teaching innovations or clarify how to implement them in the classroom (Fraser et al., 2014). The education research literature is written for education researchers and seldom provides sufficient detail or scaffolding for instructors who are not education researchers to read the methods and identify the critical elements for successful implementation (Borrego et al., 2013). PORTAAL is one effort to create a tool that translates the research-based best practices into explicit and approachable practices. It accomplishes the following:

PORTAAL identifies research-based best practices for implementing active learning in the classroom. PORTAAL distills the education literature into elements that have been shown to increase student outcomes in terms of learning, logic development, or measures correlated with both. Thus, the elements in this rubric could be considered “critical components” for the success of active learning, and fidelity to these elements should increase student outcomes as they did in the literature. In addition, the elements in PORTAAL go beyond simply identifying the amount or type of active learning occurring in the classroom and also guide instructors as to how active learning should be implemented in the classroom (with accountability, logic development, and apprehension reduction) for the greatest impact on student learning. Proper implementation of research-based best practices does not guarantee improvement in student learning, but it does increase the odds of success over less scientifically based approaches (Fraser et al., 2014).

PORTAAL deconstructs best practices into easily implemented elements. One of the reasons instructors may not attempt to change their classroom practices is because it can feel overwhelming, and no one can change their practice successfully in multiple dimensions at the same time (Kreber and Cranton, 2000). PORTAAL breaks classroom practices into four dimensions (practice, accountability, logic development, and apprehension reduction) and within these dimensions identifies discrete elements, each representing a concrete action an instructor can take to improve his or her implementation of active learning. Instructors could devote an entire year to working on one element or may decide to focus on an entire dimension. By organizing best practices into dimensions and elements, instructors now have a framework that makes the process of change more manageable.
PORTAAL provides reliable and unbiased feedback on classroom practices. Instructors traditionally get feedback on their classes through personal reflections on their own classrooms, comparing how their classes are run relative to their own experiences as a student, and student evaluations (Fraser et al., 2014). All of these methods are biased by their own interpretations or student expectations. PORTAAL aligns observable classroom practices to research-supported best practices in a reliable manner and can provide objective feedback to the instructor.

Uses of PORTAAL

Higher education currently lacks a scientific approach to teaching evaluations (Fraser et al., 2014). PORTAAL offers one such scientific approach by aligning classroom practices to research-supported best practices. PORTAAL analysis of classrooms can be useful in multiple ways:

For an instructor: Individual instructors can score their current classrooms to identify what they do well and where they could improve. Once they have identified a dimension to improve, PORTAAL offers additional elements from that dimension to incorporate into their classes. Instructors could score their classrooms over time to determine whether they have effectively changed their teaching. Instructors could include these PORTAAL scores in their teaching portfolio to document the effort they put into teaching. In addition, instructors in learning communities could observe a colleague's classroom and provide feedback based on PORTAAL observations.

For a department and college: A department could use PORTAAL to determine the level at which instructors in the department implement evidenced-based teaching practices in their classrooms and identify exemplary instructors and recognize them. They could then promote their department to incoming students as one in which instructors use best practices in STEM education. We could also imagine that an instructor's longitudinal PORTAAL scores could also be one of many measures of teaching effectiveness for tenure and promotion decisions. Colleges could use PORTAAL to document changes associated with teaching initiatives as well as documentation of teaching effectiveness for accreditation.

For researchers: Many education studies compare treatments across instructors or across classrooms. Variation between instructors in how they teach their courses could erroneously lead to conclusions about the impact of the treatment. PORTAAL observations of each classroom in a study would allow researchers to objectively compare the similarity of classrooms within or across treatment groups in terms of elements that have been shown to influence student learning. If they did not see differences, then they could be more confident that differences in outcomes were due to the treatment and not differences in classroom implementation. PORTAAL could also be used by researchers interested in faculty change, as PORTAAL could help determine how close instructors are to research-based best practices before and after a faculty development program.

In summary, active learning and evidence-based teaching practices will soon become the expected teaching method across college campuses. PORTAAL provides a suite of easy to implement and evidence-based elements that will ease instructors’ transitions to this more effective teaching method and assist instructors in objectively documenting their progress toward accomplishing this educational goal.

ACKNOWLEDGMENTS

Support for this study was provided by NSF TUES 1118890. We thank Carl Longton and Michael Mullen for help with classroom observations. We also appreciate feedback on previous drafts from Scott Freeman, Alison Crowe, Sara Brownell, Pam Pape-Lindstrom, Dan Grunspan, Hannah Jordt, Ben Higgins, Julian Avila-Pacheco, and Anne Casper. This research was done under approved IRB 38945.

REFERENCES


American Association for the Advancement of Science (2013). Describing and Measuring Undergraduate STEM Teaching Practices, Washington, DC.


Information for Authors

**CBE—Life Sciences Education (LSE)** is an online, quarterly journal owned and published by the American Society for Cell Biology (ASCB) in editorial partnership with the Genetics Society of America. The journal publishes original, previously unpublished, peer-reviewed articles on research and evaluation related to life sciences education, as well as articles about evidence-based biology instruction at all levels. The ASCB believes that biology learning encompasses diverse fields, including math, chemistry, physics, engineering, and computer science, as well as the interdisciplinary intersections of biology with these fields. One goal of the journal is to encourage teachers and instructors to view teaching and learning the way scientists view their research, as an intellectual undertaking that is informed by systematic collection, analysis, and interpretation of data related to student learning. Target audiences include those involved in education in K–12 schools, two-year colleges, four-year colleges, science centers and museums, universities, and professional schools, including graduate students and postdoctoral researchers. All published articles are available freely online without subscription. In addition, published articles are indexed in PubMed and available through PubMed Central.

**LSE** is published online four times a year: March (Spring issue), June (Summer issue), September (Fall issue), and December (Winter issue). LSE also prints a highlights issue each year in December, featuring contributions selected from the four online issues. Submissions are accepted at any time. Articles are assigned to particular issues by the editors. To be included in an issue, manuscripts must be accepted in final form at least two months prior to the publication date.

**Determining the Suitability of a Manuscript for LSE**

**Articles.** LSE is a venue for biologists to disseminate their educational innovations to others who teach biology, as well as for dissemination of biology education research that is designed to generate more generalizable, basic knowledge about biology education. Thus, LSE publishes two types of articles: descriptions of research that breaks new ground in understanding biology teaching and learning and descriptions of the implementation and evaluation of educational innovations in the life sciences. Regardless of the nature of the work, articles should offer a logical, evidence-based chain of reasoning about the design and methods used to generate the findings and support the conclusions.

The design and interpretation of studies submitted for publication in LSE should fit the goals of the work. Articles about biology education research should describe how the study was designed and conducted to yield generalizable claims and should be applicable beyond a single course or program. Authors of this type of article are encouraged to draw from the diverse social science theories, methods, and findings to inform their work, and to clearly define terms and approaches that may be unfamiliar to a biologist audience.

Articles about educational innovations should describe the systematic collection and analysis of educational data and include rigorous reflection about the results with the aim of improving instruction. Such work can be limited to a single course or program, but the educational innovation should be sufficiently novel and the results sufficiently compelling to prompt other instructors to adapt or adopt it for use with their own students. Authors of this type of article must review relevant literature to demonstrate how a particular innovation is unique compared with previously published work.

Instructors interested in publishing their educational innovations in LSE should give careful thought to how they will assess student learning or other desired outcomes. Answering three questions can help guide the process of assessment: 1) What are your instructional or programmatic goals? 2) What should learners know or be able to do if you met your goals? 3) How can you measure or otherwise document whether learners know or are able to do what you intend? Documentation of intended outcomes can be accomplished through systematic analysis of data collected through diverse approaches, such as pretest/posttest, interviews, focus groups, surveys, or performance on coursework, including exams, papers, or lab reports. Authors should present their innovations in the same way that life scientists present their research: claims regarding efficacy must be supported by evidence. Articles that lack adequate assessment, assessment instruments, descriptions of assessment methods, or references to published assessment instruments or methods will be returned to authors without review.

All articles must include collection, analysis, and interpretation of educational data, which can be quantitative or qualitative in nature. In addition, LSE articles should: 1) address a clear educational problem or education research question, 2) demonstrate clear alignment among the problem or question being addressed, the design of the study or educational innovation, the claims being made, and the evidence used to support those claims, 3) describe how results are applicable or transferable to other settings, 4) be relevant to a defined audience of educators, and 5) make reference to related educational literature. Articles should include a formal Methods section, and any assessment tools (surveys, tests, assignments, interview or focus group questions, etc.) should be included as they were administered to participants as supplemental materials. The source of the assessment tool(s) should be described, including the rationale behind the selection or design of the tool(s). The online nature of the journal facilitates the inclusion of instructional materials such as syllabi, assignments, rubrics, laboratory protocols, or professional development guidelines. Science procedures, protocols, and results that are important for understanding how instruction was accomplished should be included as supplemental materials rather than in the body of the manuscript.

The following references may be useful for thinking about the design and conduct of biology education studies:

Essays. *LSE* publishes essays on timely and important topics related to biology teaching and learning, including assessment methods, student engagement, curriculum innovations, K–20 continuum, and other topics. Essays are framed by personal experience and provide specific examples, but describe a problem or approach of general interest and may be synthetic across the work of many individuals. Appropriate foci for essays include reviews of current practices, policies, or research that have implications for biology teaching and learning, or personal perspectives on issues that are provocative or would otherwise be of widespread interest. The problem or approach should be presented within a scholarly context, citing references and resources that address the topic. If claims are to be made, there should be evidence from the literature or the authors' own work. Although it is not a requirement, essays can include ideas for assessment or future research as appropriate. Manuscripts that include claims about the efficacy of an instructional approach should be submitted as articles. Descriptions of studies with preliminary or very limited data will not be considered.

*LSE* also publishes Research Methods essays that offer scholarly and practical advice on biology education research design and methods. Research Methods essays should be written as instructional pieces, identifying common and significant methodological issues. These essays should focus on a single topic, treating it with sufficient depth for readers to understand and be able to take action on the issue, while appealing to a broad audience of biology education scholars and education-interested biologists. Authors are encouraged to be creative in format. For example, Research Methods essays could be scholarly reviews punctuated by practical advice, or in-depth discussions of articles that illustrate exemplary methodological practice. The essays should be concise and accurate yet approachable, clearly defining technical terms and using biology-friendly analogies and examples to illustrate key points. Co-authorship by social scientists and biologists is encouraged.

Meeting Reports. The journal publishes occasional reports of meetings related to education in the life sciences. Such reports should not merely describe a meeting, but should contain material that can be immediately used by the reader. Examples include, but are not limited to, the following: announcement and description of an ongoing dialogue with an invitation and ways to participate; report of new findings described at the meeting with access to the underlying materials via citations or web resources; report of new findings based on the meeting dialogue *per se*; compilation and synthetic description of accessible resources described or discussed at the meeting, complemented by clear recommendations for action or next steps; creation of new opportunities to participate in planned workshops, online dialogues, or the like. Meeting reports received through the normal online submission process will be judged on the basis of their relevance and interest for the broad readership of the journal. Prospective authors are encouraged to contact the editors in advance to discuss the suitability of a given meeting report.

Letters to the Editor. A goal of *LSE* is to stimulate dialogue. *LSE* invites readers to make use of the Reader Comment function associated with each publication or submit Letters to the Editor. *LSE* requires that letters are directly responsive to an article published in the journal or that they bring a new or under-recognized issue to the attention of readers in a way that is informed by the inclusion of relevant data or references to the literature. Letters received through the normal online submission process will be reviewed by the Editor-in-Chief and published at her discretion. Appropriate topics for letters include commentary on education-related articles in this or other journals as well as opinions on more general topics of concern to the readership.

Features. Articles listed under the heading of “Features” are by invitation only. Authors interested in contributing to a feature should contact the editor.

Announcements. *LSE* accepts noncommercial announcements of meetings, workshops, and conferences and of funding opportunities and fellowships open to all.

Peer-Review Process

All submitted manuscripts and educational materials are subject to peer review. After a manuscript has been submitted to *LSE*, the Editor-in-Chief selects an editorial board member to guide the paper through the review process. Editorial board members select two reviewers to submit written evaluations. The board member will assess the peer reviews and determine whether the submission will be accepted as is, accepted with suggested revisions, temporarily rejected with suggestions for improvements before resubmission, or rejected with reasons explaining this decision. The corresponding author can usually expect an initial response within four weeks.

License and Publishing Agreement

Authors are required to sign a License and Publishing Agreement when a manuscript is accepted for publication. Under the terms of that agreement, authors retain copyright but grant the ASCB a perpetual license to publish the manuscript. Authors also grant to the general public the nonexclusive right to copy, distribute, or display the manuscript subject to the terms of the Creative Commons—Noncommercial—Share Alike 3.0 Unported license (http://creativecommons.org/licenses/by-nc-sa/3.0/).

ASCB Policy on Research Misconduct by Authors

By submitting a paper to *LSE*, an author acknowledges that he or she is subject to the ASCB Policy on Research Misconduct by Authors. The policy is posted at www.ascb.org/files/research_misconduct.pdf.
Institutional Review Board Approval

Manuscripts reporting on studies involving human subjects must include explicit assurance that the research reported was approved by a local Institutional Review Board, including an Institutional Review Board number, unless the research is exempt from such review according to U.S. Department of Education guidelines. Prospective authors are advised that permission must be obtained in advance.

Guidelines for Preparing Articles, Essays, and Features

General Instructions. Submitted papers must be original (i.e., not published or submitted for publication elsewhere). Previously published material will not be considered. Publication of a short abstract does not constitute prior publication, nor does presentation of data at a scientific meeting or in a webcast of such a meeting. The posting of a manuscript on an author’s personal website or in an online institutional repository prior to submission of the manuscript to LSE is not considered prior publication.

The text should be written in clear, concise, and grammatical English. Manuscripts ordinarily begin with an overview of how the work presented is relevant to the classroom, laboratory, or curriculum and what student outcomes are expected. Whenever possible, incorporate materials by citing relevant publications, without repeating already published works. Manuscript files must be submitted in .doc, .docx, or .rtf format.

Tables and illustrations should convey information effectively and must be uploaded separately. Graphs and figures should be provided digitally as separate TIF or EPS files. Authors are encouraged to take advantage of the online nature of LSE. Video, audio, databases, images, animations, molecular structures, and other electronic resources may be linked as supplemental material for further consideration by readers.

International authors may wish to consider using an editorial service, such as ScienceDocs (www.sciencedocs.com), The Medical Editor (www.themedicateditor.com), American Journal Experts (www.jou rnalexperts.com), Editage (www.editage.com), Bioscience Writers (www.biosciencewriters.com), Cognyte Ltd. (www.cognyte.co.uk), or Squirrel Scribe (www.squirrelscribe.com). LSE does not endorse any particular editorial service, such as ScienceDocs (www.sciencedocs.com), The Medical Editor (www.themedicateditor.com), American Journal Experts (www.jou rnalexperts.com), Editage (www.editage.com), Bioscience Writers (www.biosciencewriters.com), Cognyte Ltd. (www.cognyte.co.uk), or Squirrel Scribe (www.squirrels crite.com). LSE does not endorse any particular editorial service.

Institutional Review Board Approval

Manuscripts reporting on studies involving human subjects must include explicit assurance that the research reported was approved by a local Institutional Review Board, including an Institutional Review Board number, unless the research is exempt from such review according to U.S. Department of Education guidelines. Prospective authors are advised that permission must be obtained in advance.

Length Guidelines. The following manuscript submission lengths are intended to aid authors in preparing their manuscripts; however, submissions outside these ranges will be considered.

Articles: 30,000–60,000 characters (with spaces), or 5–10 journal “pages”; typically do not exceed 20 journal pages, or 120,000 characters.

Essays: 12,000–30,000 characters (with spaces), or 2–5 journal “pages”; typically do not exceed 10 journal pages, or 60,000 characters.

Features: 6,000–12,000 characters (with spaces), or 1–2 journal “pages.”

Letters: 3,000–6,000 characters (with spaces), or up to half a journal “page.”

Cover Letter. Authors should submit a cover letter from the corresponding author stating that the work is being submitted exclusively to LSE and indicating why it is appropriate for the journal. If there is a connection between an author and a commercial product being used or reported, full disclosure is required in the cover letter and appropriate statements should be included in the manuscript. (See “Title Page” below.) Authors are invited to suggest monitoring editors and reviewers (please include institution and email address). For article submissions, authors are encouraged to indicate whether their work is best described as research that aims to yield new insights about biology teaching and learning or as the implementation and evaluation of educational innovations in the life sciences.

Title Page. Page 1 should include the title of the manuscript, the type of manuscript being submitted (e.g., article, essay, feature, letter to the editor, response), the number of characters in the manuscript, a shortened running title (not to exceed 42 characters and spaces), and the names and affiliations (including department, institution, city, state, and zip code) of all authors in the order in which they should appear. List the corresponding author separately with complete postal and email address and telephone and fax numbers. Keywords should also appear on page 1. Include at least five keywords selected from the text of the article. If possible, include a keyword that indicates the target learners (primary, secondary, undergraduate, graduate, general public, etc.).

If one or more of the authors of a research paper that assesses the effectiveness of a product or curriculum was also involved in producing the product or curriculum, readers need to be fully aware of this potential conflict of interest. Therefore, any potential conflicts of interest should be clearly stated on the title page of the manuscript. The author and the product should be identified, and a statement included that no promotion of a particular product to the exclusion of other similar products should be construed. This will be noted under the byline if the manuscript is accepted for publication.

Abstract. Page 2 should contain the abstract, which should be no more than 200 words long and should summarize the important points in the manuscript.

Manuscript Text. The text of the paper should begin on page 3. LSE follows the style guidelines of the Council of Biology Editors Style Manual. For chemical nomenclature, follow the Subject Index of Chemical Abstracts. Capitalize trade names and give manufacturer names and addresses. Do not include figures or tables within the body of the manuscript. A format of Introduction, Methods, Results, Discussion, and References is encouraged, but other formats may be more appropriate for some topics. Manuscripts should include line and page numbers.

Accessing Materials. Describe how to access new educational materials if the study or use of such materials is the subject of the paper. If materials are online, provide a URL to the material. Any registration requirements or agreements inherent in the use of the materials should be described. If there are no online materials, simply state “No additional materials available online.” For other new educational materials presented in the manuscript, authors should describe how readers can access the materials, what format is
Acknowledgments. Identify financial sources and other sources of support for the research being reported in the manuscript.

References. Place the reference list immediately following the manuscript text (beginning on a new page). Consult the most recent issue of LSE for reference citation formats. Cite references in the text alphabetically by name and date (e.g., Jones and Smith, 1987; Smith et al., 1988). Only published articles or manuscripts accepted for publication can be listed in the Reference section. References should contain complete titles and inclusive page numbers and should be listed in alphabetical order. Abbreviate the names of journals as in PubMed. The abbreviations can be found at LinkOut Journal Lists. Reference citations may include up to 10 authors followed by et al.

Citations to online websites are acceptable reference links. In the text, include the name of author for the cited website and the date accessed (e.g., National Center for Biotechnology Information, 2005). The citation should include the author (if known), date of latest modification (if known) or copyright, title of page, title of complete work (if appropriate), URL (e.g., www.website.org), and date of access. For example: National Center for Biotechnology Information (2005). NCBI Home Page. www.ncbi.org (accessed 5 February 2005).

Unpublished results, including personal communications and submitted manuscripts, should be cited as such in the text. Personal communications must be accompanied by permission letters unless they are from the authors’ own work.

Footnotes. Call out footnotes at the appropriate place in the text with a superscript numeral. The footnote text should be placed on a separate page after the References.

Figures. All figures should be uploaded as individual files.

Figure Legends. Figure legends should appear in numerical order after the References. Figure legends should provide an overview of the figure and details that describe any component parts.

Tables. All tables must be cited in order in the text of the manuscript. Individual table files need to be uploaded separately.

Supplemental Material. Upload all supplemental material (except in the case of videos) together in one combined PDF. Be sure to upload the final version of the supplemental material. It will be posted online as received and will not be edited.

Guidelines for Preparing Digital Artwork

Digital artwork must accompany the manuscript submission. Figures should be uploaded as separate files with the manuscript through the online manuscript submission system. Because artwork must be of sufficient quality for print reproduction, LSE asks that all artwork be prepared using professional graphic art software. Word processing and presentation software packages (such as Word and PowerPoint) are inadequate for preparing high-quality digital artwork.

Prepare all digital artwork as RGB TIF images, at 300 dpi resolution, or EPS images:

Figure Size. Prepare figures at the size they are to be published.
Up to 1 column wide: Figure width should be 4.23–8.47 cm.
1 to 1.5 columns wide: Figure width should be 10.16–11.43 cm.
2 columns wide: Figure width should be 14.39–17.57 cm.
The figure depth must be less than or equal to 23.5 cm.

Guidelines for Preparing Electronic Resources

It is possible for authors to submit for peer review electronic works including, but not limited to, animations, Chime tutorials, movies, interactive websites that may include quizzes, images (electron micrographs, photomicrographs, etc.), Java Applets, searchable databases, etc. Articles that describe new educational uses of existing resources (e.g., Expression Connection, FlyBase, Database of Interacting Proteins, etc.) are also of interest. Manuscripts of this type should provide detailed instructions for use of the resource by the target audience.

For works using someone else’s electronic resource (such as a database), a letter from the creator or curator of the resource indicating their willingness to support free pedagogical use of their work must be included. The same general rules for evaluation will apply to all electronic submissions. All submissions will be evaluated for (1) pedagogical content, (2) clear description of goals and expected student outcomes, (3) transferability to other settings, (4) appropriateness for the target audience, and (5) references to related educational literature. After publication of the electronic work(s), the authors will be encouraged to submit the work to other databases (National Digital Library, BEN, etc.); however, it is expected that the LSE publication citation will remain associated with the work. This will allow viewers to read a more in-depth discussion of the work.

All such electronic works must be freely available, and will be hosted on the LSE server or on the ASCB server, with the exception of large databases. This will ensure stable access to the works with a nonchanging URL.

A manuscript should accompany any electronic submission. The manuscript should describe 1) the learning goals or purpose of the electronic work, 2) the target audience, 3) development of the electronic work (describe hardware and software used), 4) platform availability (see below), 5) a description of any necessary hardware or software, with links to the appropriate sites for downloading (e.g., plugins, helper applications, etc.), and 6) assessment of the work’s impact on student learning.
Ideally, submitted works should work on any platform (PC, Mac, Unix) and on all browsers. If there are known restrictions, these should be included in the manuscript. LSE can help authors test their works for such limitations if they do not have access to certain platforms or browsers.

The electronic work may have been hosted previously on any website, but the authors may not have previously published any description of the electronic work other than the associated Web pages. Published journal descriptions of the electronic work will preclude publication in LSE, with the exception of abstracts or presentations at professional meetings.

The online publication will include hyperlinks to the work that will appear in a new browser window, if appropriate. This capacity could be helpful to the authors since they could provide directions for readers as needed to illustrate particular aspects of the work. The layout and submission process for the manuscript accompanying an electronic work should follow the same general format as other categories. The electronic work should be submitted to LSE at the same time as the manuscript. If this presents a problem, contact LSE staff for assistance.

Sharing Materials and Data
Publication of a manuscript in LSE implies that the authors agree to make available, to the extent legally permissible, all propagative materials such as mutant organisms, cell lines, recombinant plasmids, vectors, viruses, monoclonal antibodies, instructional materials, and assessment instruments that were used to obtain results presented in the article. Prior to obtaining these materials, interested scientists will provide the authors with a written statement that they will be used for noncommercial research purposes only. Authors are encouraged to share raw data with qualified researchers who wish to reproduce or further analyze the authors’ work for noncommercial purposes. Sharing of data on human subjects should be consistent with the conditions of the Human Subjects protocol approved by the authors’ Institutional Review Board (and any other agreements made with the subjects) for the work reported.

How to Submit Manuscripts
Electronic Submission. Authors must submit manuscripts online at www.cellbiologyeducation.org. Specific instructions on how to submit your manuscript are available at the submission site. Authors should submit the manuscript in .doc, .docx, or .rtf as two text files, the first containing the cover letter and the second containing the manuscript file. Figures, tables, and Supplemental Material, including Appendixes, must be submitted as individual files. Do not embed figures in the manuscript. If you are submitting a feature that does not contain an abstract or keywords, write “There is no abstract” in the required abstract field and “one,” “two,” and “three” in the keyword fields.

Questions regarding submission guidelines can be directed to: cbe@ascb.org or 301-347-9304.

General Questions
At any stage in the submission process, authors with questions should contact the LSE Editorial Office at 301-347-9304 (phone); 301-347-9350 (fax); cbe@ascb.org; or The American Society for Cell Biology, 8120 Woodmont Avenue, Suite 750, Bethesda, MD 20814-2762.

You are encouraged to contact editorial board members or the Editor-in-Chief by email to discuss submissions to LSE.

The LSE website is an additional resource to authors. See earlier issues of LSE for examples of the different types of manuscripts published.

Last updated on 4/13/2015.