2016 HIGHLIGHTS

Table of Contents

FEATURES
WWW.Life Sciences Education
Bringing Climate Change into the Life Science Classroom: Essentials, Impacts on Life, and Addressing Misconceptions
Amy J. Hawkins and Louisa A. Stark ................................................................. 1–6

Most life science educators are unaware of the scientific consensus regarding climate change. This Feature highlights free online resources to teach climate change, including research summaries, materials about the life science impacts of climate change, and resources that address common misconceptions and “skeptical” arguments.

From the National Academies
Insights from a Convocation: Integrating Discovery-Based Research into the Undergraduate Curriculum
Sarah C. R. Elgin, Gita Bangera, Sean M. Decatur, Erin L. Dolan, Laura Guertin, Wendy C. Newstetter, Elyvra F. San Juan, Mary A. Smith, Gabriela C. Weaver, Susan R. Wessler, Kerry A. Brenner, and Jay B. Labov ................................................................. 7–13

The National Academies of Sciences, Engineering, and Medicine organized a convocation in 2015 to explore and elucidate opportunities, barriers, and realities of course-based undergraduate research experiences, known as CUREs, as a potentially integral component of undergraduate science, technology, engineering, and mathematics education. This paper summarizes the convocation and resulting report.

ESSAY
A Conceptual Framework for Graduate Teaching Assistant Professional Development Evaluation and Research
Todd D. Reeves, Gili Marbach-Ad, Kristen R. Miller, Judith Ridgway, Grant E. Gardner, Elisabeth E. Schussler, and E. William Wischusen ................................................................. 15–23

This essay proposes a conceptual framework for biology graduate teaching assistant professional development evaluation and research with three overarching variables for consideration: outcome variables, contextual variables, and moderating variables.

RESEARCH METHODS
A Primer for Developing Measures of Science Content Knowledge for Small-Scale Research and Instructional Use
Kristin M. Bass, Dina Drits-Esser, and Louisa A. Stark ................................................................. 25–38

This essay, intended for faculty involved in small-scale projects, courses, or educational research, provides a step-by-step guide to the process of developing, scoring, and validating content knowledge assessments. The authors illustrate their discussion with examples from their measures of high school students’ understanding of cell biology and epigenetics.
Contemporary Test Validity in Theory and Practice: A Primer for Discipline-Based Education Researchers

Todd D. Reeves and Gili Marbach-Ad .................................................. 39–47

This essay offers a contemporary social science perspective on test validity and the validation process. The instructional piece explores the concepts of test validity, the validation process, validity evidence, and key threats to validity. The essay also includes an in-depth example of a validity argument and validation approach for a test of student argument analysis. In addition to discipline-based education researchers, this essay should benefit practitioners (e.g., lab directors, faculty members) in the development, evaluation, and/or selection of instruments for their work assessing students or evaluating pedagogical innovations.

ARTICLES

Career Development among American Biomedical Postdocs
Kenneth D. Gibbs, Jr., John McGready, and Kimberly Griffin ................................. 49–60

This study reports results from a national survey examining the career development of biomedical postdocs. Findings point to the need for enhanced career development programs earlier in the training process and interventions that are sensitive to distinctive patterns of interest development across social identity groups.

Cues Matter: Learning Assistants Influence Introductory Biology Student Interactions during Clicker-Question Discussions
Jennifer K. Knight, Sarah B. Wise, Jeremy Rentsch, and Erin M. Furtak ......................... 61–74

Recordings of introductory biology student discussions of clicker questions demonstrate that students use reasoning and questioning in their discussions and that their use of these discussion characteristics is heavily influenced by the cues they hear from learning assistants during discussions.

Development and Assessment of Modules to Integrate Quantitative Skills in Introductory Biology Courses
Kathleen Hoffman, Sarah Leupen, Kathy Dowell, Kerrie Kephart, and Jeff Leips .................. 75–86

Recent reports have called for integrating quantitative reasoning into biology courses to prepare students for careers in research and medicine. This article describes the development, implementation, and assessment of modules designed to incorporate quantitative reasoning and skill development into introductory biology courses.

Features of Knowledge Building in Biology: Understanding Undergraduate Students’ Ideas about Molecular Mechanisms
Katelyn Southard, Tyler Wince, Shanice Meddleton, and Molly S. Bolger ......................... 87–102

This article explores how undergraduate students integrate knowledge in molecular biology. Analysis of interviews with introductory-level and upper-division students revealed patterns in how students sorted and connected ideas about DNA transcription, translation, and replication. Findings include differences in the nature of how students connected ideas.
Scientist Spotlight Homework Assignments Shift Students’ Stereotypes of Scientists and Enhance Science Identity in a Diverse Introductory Science Class
Jeffrey N. Schinske, Heather Perkins, Amanda Snyder, and Mary Wyer ..........................................................103–120

The authors evaluated a series of homework assignments featuring counterstereotypical examples of scientists in an introductory biology class. Following the intervention, students exhibited nonstereotypical views of scientists and conveyed an enhanced ability to personally relate to scientists. These shifts correlated with science interest and course grades.

New Measures Assessing Predictors of Academic Persistence for Historically Underrepresented Racial/Ethnic Undergraduates in Science
Angela Byars-Winston, Jenna Rogers, Janet Branchaw, Christine Pribbenow, Ryan Hanke, and Christine Pfund ..........................................................121–131

Using social cognitive career theory and science identity theory, the authors validate new measures assessing persistence-related factors with students from historically underrepresented racial/ethnic groups in science.

Early Engagement in Course-Based Research Increases Graduation Rates and Completion of Science, Engineering, and Mathematics Degrees
Stacia E. Rodenbusch, Paul R. Hernandez, Sarah L. Simmons, and Erin L. Dolan ..............................................133–142

This study shows that participation in course-based undergraduate research experiences (CUREs) improves students’ likelihood of graduating with a STEM degree and graduating within 6 years. These results support offering CUREs in place of standard lab courses as an effective strategy for producing additional college graduates with STEM degrees.

Colleagues as Change Agents: How Department Networks and Opinion Leaders Influence Teaching at a Single Research University
T. C. Andrews, E. P. Conaway, J. Zhao, and E. L. Dolan ..........................................................143–159

This study investigates colleague–colleague relationships within life sciences departments and how they may promote teaching reform. The authors found that discipline-based education researchers are perceived as promoting changes among colleagues in views about teaching and teaching practices. Other faculty may also be leveraged to support change.

A Campus-Wide Investigation of Clicker Implementation: The Status of Peer Discussion in STEM Classes
Justin D. Lewin, Erin L. Vinson, MacKenzie R. Stetzer, and Michelle K. Smith ..............................................161–172

The authors observed university STEM classes and documented clicker use. Comparisons of classes taught with and without clickers show that the use of clickers does not significantly impact lecture time. One explanation stems from the observation of three distinct modes of clicker use that differ in instructional behaviors and question difficulty.

Information for Authors ..........................................................173–177
Bringing Climate Change into the Life Science Classroom: Essentials, Impacts on Life, and Addressing Misconceptions

Amy J. Hawkins* and Louisa A. Stark

Genetic Science Learning Center, University of Utah, Salt Lake City, UT 84102

Climate change is at the forefront of our cultural conversation about science, influencing everything from presidential debates to Leonardo DiCaprio’s 2016 Oscar acceptance speech. The topic is becoming increasingly socially and scientifically relevant but is no closer to being resolved.

Most high school students take a life science course but significantly fewer take earth or environmental science. This leaves life science educators with the critical task of conveying information about climate change despite many having never received any formal education on the topic themselves. Earlier this year, Science magazine and the National Center for Science Education published results from the first nationally representative survey on how teachers approach climate change in their classrooms (Plutzer et al., 2016a, b). The researchers found that most middle school science teachers and high school biology teachers allocate at least an hour to discussing recent global warming. However, most of these teachers were not aware of the extent of scientific consensus on the topic—97% of climate scientists agree that climate change is caused by human activity. This gap in knowledge may create a sense of pressure for teachers to provide a “balanced” presentation of the issue, although the evidence suggests that to do so would only serve to perpetuate misconceptions and erode the validity of scientific consensus in the minds of students. Thus, even high-achieving students may arrive at college with gaps in their knowledge that need to be addressed before moving on to study the consequences of climate change within the life sciences.

In this Feature, we highlight free online resources that will be useful for science educators teaching climate change to high school and undergraduate students.

PRIMERS AND REINTRODUCTIONS

For teachers looking for a quick and thorough overview of the most basic and compelling data on climate change, an excellent starting place is the introductory section “Current Climate Change” in the Climate Science Primer available through the U.S. Forest Service Climate Change Resource Center (www.fs.usda.gov/ccrc/climate-basics/climate-primer). The primer is text-based and concise and includes an animation showing the relationship between temperature and atmospheric CO₂ from 800,000 yr ago to the present (1:31 min). The site also includes several data-rich graphs: the rise in temperature from 1880 to the present, the correlation between temperature and atmospheric CO₂ over the past 450,000 yr, the rise in sea level since 1870, and the decline in Antarctic ice mass.

Subsequent sections of the primer go on to explain the mechanism of global warming—the greenhouse effect, human influence on the greenhouse effect, natural climate cycles, and the effects of climate change on ecosystems in the United States, including modeling predictions of changes in precipitation and temperature. The Climate Change Education portion of the website features two Education Modules (www.fs.usda.gov/ccrc/climate-basics/education; Figure 1). Climate Change Science and Modeling (22:57 min), and Climate Change Effects on Forests and Grasslands (20:32 min). The modules are narrated video interactivesthatahref="http://climate.nasa.gov/"target="_blank" that have the option for closed-captioning. Text versions of the narration are also available. Each section of the Climate Change Resource Center is well referenced with peer-reviewed research articles.

The National Aeronautics and Space Administration’s (NASA) Global Climate Change: Vital Signs of the Planet website (http://climate.nasa.gov) provides extensive and stunning graphics and multimedia resources to visualize climate data. Students can access short videos featuring satellite imagery and images from space, infographics on the rise in global temperatures and sea level, and quizzes on the topics covered on the website (http://climate.nasa.gov/resources/graphics-and-multimedia/; Figure 2). The website also offers several mobile apps for Apple devices: Earth Now, which displays real-time global satellite data (also available for Android devices); Images of Change, a collection of photographs and satellite data that show the effects
of climate change over time; and Offset, a retro-style educational game about slowing the effects of global warming through the use of alternative energy sources (http://climate.nasa.gov/earth-apps/). While relatively few of the materials explicitly incorporate life science, the visual quality of the materials alone makes the website worth visiting.

Climate Literacy: The Essential Principles of Climate Science (2009; 18 pp.) is a text-based interagency guide published by the U.S. Global Change Research Program. This guide provides a framework and seven essential principles for formal and informal education about climate change (http://downloads.globalchange.gov/Literacy/climate_literacy_highres_english.pdf). It is available in English and Spanish. For those seeking further non English language resources, the U.S. Environmental Protection Agency has translated their website, including a section on climate change, into five languages (www3.epa.gov).

**IMPARTS OF CLIMATE CHANGE ON NONHUMAN LIFE**

Although a multitude of resources exist on how and why climate change is occurring, there are relatively few resources that provide details about the impacts of climate change on plants and animals. Ecological Impacts of Climate Change (2009; 32 pp.) from the National Academies presents examples of impacts on specific plant and animal species in each of seven regions of the United States—the Pacific coastline, Alaska and the Arctic, the western mountains, the southwestern deserts, and the central, southeast, and northeast regions. The booklet is abundantly illustrated with photographs and other visuals and would be appropriate to assign as an introductory, out-of-class reading on climate change and ecology. It is freely available as a downloadable PDF at http://oceanservice.noaa.gov/education/pd/climate/teachingclimate/ecological_impacts_of_climate_change.pdf.

The Quest website from KQED Science provides several resources that can be found in the Topics (Climate) and Collections (Ocean Acidification) drop-down menus under the website title (ww2.kqed.org/quest). For example, the Giant Sequoias Struggle with Drought movie (6:58 min; ww2.kqed.org/quest/2015/11/17/giant-sequoias-struggle-with-drought/) presents the effects of the warming climate and associated drought on tree species, including sequoias. Researchers show how they collect samples from the canopies of these ancient, tall trees and use a pressure chamber to measure the trees’ levels of water stress. An animation that shows how water stress can lead to tree death also makes this movie appropriate for plant anatomy or physiology classes. Napa Wineries Face Global Warming (7:20 min; ww2.kqed.org/quest/2015/01/06/napa-wineries-face-global-warming/) describes how temperature changes will lead to unpredictable changes in the areas of the United States that are appropriate for growing particular varieties of grapes. The section discussing the impact of temperature on sugar and anthocyanin production in grapes, and thus wine, makes this movie particularly appropriate for plant biology or economic botany classes. Balancing Act: Otters, Urchins and Kelp (7:19 min; ww2.kqed.org/quest/2014/02/25/balancing-act-otters-urchins-and-kelp/) illustrates the complexity of natural ecosystems and their impact on CO2 levels. The reintroduction of sea otters on the west coast of Washington led to a reduction in sea urchins (which the otters eat) and an increase in kelp (on which the urchins graze). Kelps are one of the most efficient absorbers of atmospheric CO2; they can absorb up to 12 times more CO2 when sea otters are present. Ocean Acidification and Marine Life (ww2.kqed.org/quest/2014/12/12/ocean-acidification-and-marine-life) is an animated set of...
seven slides that provides an introduction to the chemistry of ocean acidification and its effect on shell formation.

A free iBook, Clue into Climate, also from KQED, includes Chapter 3, “Changing Ecosystems,” with several embedded movies. A Warmer World for Arctic Animals (3:02 min) describes the impacts of warming temperatures on polar bears, walruses, arctic foxes, and Beluga whales. The California at the Tipping Point: Intertidal Zone excerpt (4:14 min) discusses the impact of warmer ocean temperatures and rising ocean levels on tidal pool ecosystems. Some Bugs Like It Hot (10:12 min) describes how warmer temperatures have led to a large increase in the overwintering range of the potato–tomato psyllid, leading to significant crop damage and decreased income for farmers. Another economic botany movie in Chapter 1, “Causes of Change,” describes the impacts of warmer weather on the cherry crop—Uncool Cherries (6:37 min). All of these movies have high production values and feature scientists at several career stages. The iBook chapters also include simple interactives that address topics related to climate change. Some of the interactives are also available in the “Clue into Climate” collection on the Quest website.

Earth Vision Institute’s Getting the Picture: Our Changing Climate online multimedia tool (http://gettingthepicture.info/; Figure 3) provides an introduction to climate change. Section 4, “Our Changing Biosphere,” includes examples of the effects of climate change on phytoplankton, penguins, polar bears, coral reefs, and Rocky Mountain forests and discusses phenology (the study of cyclical and seasonal natural phenomena) and the impact of changing seasons on pollinators. The text is supported by short movies, animations, enlargeable photos, graphics, and a clickable U.S. map describing the responses of 30 organisms to climate change.

The National Park Service (NPS) has produced a series of movies (3:05–8:58 min) on the consequences of climate change for plants and animals in national parks across the United States. Unfortunately, there does not seem to be a single location with links to all of the movies. We found some on the NPS Climate Change website (www.nps.gov/subjects/climatechange/sciencevideos.htm). For example, Ocean Acidification at Point Reyes National Seashore (7:43 min) presents the effects of ocean acidification on multiple organisms in the ecosystem and on food webs. It features primarily women scientists, who provide background information and discuss several research studies. Species Range Shifts at Sequoia-Kings Canyon National Park (8:530 min) is unusual in that it raises questions about the choices we will have to make about species preservation as the climate changes. We found movies from additional parks by searching “climate change” on the NPS website (www.nps.gov). The NPS movies provide local information for those in the vicinity of these parks.

**Databases of searchable resources**

Several websites curate lists of hundreds or thousands of multimedia resources that can be searched using multiple filters, including intended audience, resource type, and scientific subheadings. Though these websites will likely deliver overlapping results, the different foci of the parent organizations and searchable filters allow educators to discover resources that can be more exactly matched to fit their curricular needs. While all of these resources are relevant for teaching climate change, a minority of them have an explicit focus on life science. Most of them focus on CO₂ concentrations, global temperature models, or sea-level rise.

The Collection of Climate and Energy Educational Resources Portal (CLEAN; http://cleanet.org) is one of the most extensive collections and lists 655 resources that are sorted as activities, short demonstrations/experiments, visualizations, or videos. These resources are continually updated by qualified scientists and educators from institutions that include the National Oceanic and Atmospheric Administration, the Science Education Resource Center at Carleton College (SERC), the Cooperative Institute for Research in Environmental Sciences, and TERC (Technical Education Research Centers). These professionals engage in a rigorous review process, examining each resource for its accuracy and alignment with Climate Literacy and Energy Literacy frameworks and the Next Generation Science Standards.

The CLEAN visual resources range in complexity. For example, visitors can find single static images, including a map of the global Climate Demography Vulnerability Index showing human populations with the highest vulnerability to climate change (www.mcgill.ca/files/_nea/172326_GlobalVulnerabilitymap.pdf).

An extensive, interactive website, Global Forest Watch, hosts a forest-monitoring tool that permits users to make country-by-country or global comparisons of tree cover loss and other ecological or geopolitical information (www.globalforestwatch.org/map/; Figure 4). Beyond presenting complicated data in several interactive formats, Global Forest Watch is notable in that its interactive map and supporting blogs can be viewed in 15 different languages.

The CLEAN Collection overlaps with the 3548 resources listed at SERC when searching for “climate change” (http://serc.carleton.edu/serc/search.html). In addition to its searchable database, SERC also has a site guide specific...
for climate change and global warming (http://serc.carleton.edu/serc/site_guides/climate.html), which allows users to browse 707 climate change activities and 262 climate change courses (65 and 17 of which are cross-listed with biology in their search engine, respectively). Many of these resources are text-based classroom activities or syllabi. For example, the “Malaria and Global Warming” exercise was developed for an introductory college-level biology course by Dale Blum at Pierce College (http://serc.carleton.edu/bioregion/examples/59109.html). In it, students are asked to work in small groups and map where they predict malaria will be found in the next 50 yr based on information about factors that affect the distribution of malaria in North and South America.

National Geographic published a special issue on climate change and offers 75 teacher resources appropriate for high school or postsecondary students (http://education.nationalgeographic.org/topics/climate-change/). Like the previous databases, these resources can be sorted by audience and format. An example of a relevant high school activity designed to take 45 min is “Sources, Sinks, and Feedbacks: What Role do the Oceans Play in Global Warming?” (http://education.nationalgeographic.org/activity/sources-sinks-and-feedbacks/). This activity lets students explore the relationships between ocean surface temperature and levels of atmospheric CO₂ and water vapor.

The National Science Digital Library connects to 1978 educational resources on climate change (https://nsdl.oercommons.org/search?f.search=climate+change); of these, 700 are listed as being appropriate for high school or college audiences and are cross-listed with the subject area “life science.” In addition to being able to search by subject area, education level, and material type, users can refine their searches by including education standards; results can be sorted by title, relevance, rating, or date. A quick search turned up “AH-CHOO! A Case Study on Climate Change and Allergies” (http://nsdl.oercommons.org/courses/ah-choo-a-case-study-on-climate-change-and-allergies/view), which is hosted by the National Center for Case Study Teaching in Science. In this text-based activity, students engage in scientific argumentation by role-playing a public relations specialist to evaluate data and develop materials to communicate the link between climate change and pollen allergies.
ADDRESSING MISCONCEPTIONS AND CLIMATE CHANGE DENIAL

Perhaps the most fundamental barrier to teaching and communicating about climate change is the fact that some people treat the topic as controversial despite a clear consensus in the scientific community about its causes and effects. As such, educators may find it helpful to introduce the topic by first addressing misconceptions and deliberate misinformation perpetrated about climate change. This approach is not without its hazards, as detailed in The Debunking Handbook, a nine-page document that is one of many clear and well-referenced resources available through the website Skeptical Science (www.skepticalscience.com/Debunking-Handbook -now-freely-available-download.html).

A common approach to dispelling misconceptions rests on the assumption that they exist due to a lack of knowledge and that simply supplying more information will displace the misinformation. This communication strategy, known as the information deficit model, tends to fail when directed at deeply entrenched ideas, however tenuously supported by evidence or wishful thinking. Several “backfire effects” have been identified in communication research. These show that attempts to refute misinformation can actually strengthen people’s incorrectly held beliefs by increasing their familiarity with the myths themselves, by providing too many counterarguments that make the original myth attractive in its relative simplicity, or by attempting to dispel myths that are tied to a cultural identity or worldview that is linked to a person’s sense of self.

Skeptical Science (www.skepticalscience.com) provides summaries of current peer-reviewed climate science organized around the claims most frequently made by climate change “skeptics.” The website’s user interface is a bit cluttered but includes resources that have been translated into more than a dozen languages. The information is updated frequently with original material and cross-posts from other websites, including text, videos, and graphical representations of data, that deal with either climate science or climate science politics. The website also offers free smartphone apps for Apple, Android, or Nokia devices that allow users to view the most commonly used skeptic arguments coupled with summaries and links to the original scientific research that refutes each argument (Figure 5). Skeptical Science also has made freely available their self-paced Massive Open Online Course (MOOC) on Denial101x: Making Sense of Climate Science Denial. The course materials—in tended to serve as resources for teachers—contain 81 short lectures, each ~5–7 min long, and 40 full interviews with climate science experts (www.skepticalscience.com/denial101x -videos-and-references.html).

Two other video-driven websites that offer resources to deal with misconceptions include the science video blog/YouTube channel Veritasium (www.youtube.com/user/1veritasium) and ClimateCrocks.com. Derek Muller from Veritasium has created an entertaining and referenced video on 13 Misconceptions about Global Warming (6:49 min; published on September 22, 2014; www.youtube.com/watch?v =OWXoR51xyIU). While it is unlikely to be universally appreciated, his irreverent tone may be particularly appealing to high school and college students. His video Climate Change Is Boring (4:55 min; published on August 29, 2014) may also be appropriate as an introductory assignment to show that, while the public might conceptually link climate change with the hole in the ozone level or the individual decision to engage in personal consumer recycling, the main driver of climate change is greenhouse gas emissions (www.youtube.com/watch?v=ENx9tvCrvv8).

“Inoculating” one’s audience by explaining the techniques used to spread climate change misinformation has been shown to strengthen resistance to unscientific claims by bolstering critical-thinking skills. Peter Sinclair’s frequently-updated blog ClimateCrocks.com (http://climatecrocks.com) sets out to accomplish this by responding to media coverage of climate change. The blog spans from July 2009 to the present. Its short video series, Climate Denial Crock of the Week, features excerpts from interviews with climate and policy experts, speaking at a level appropriate for the educated layperson (http://climatecrocks.com/category/crock-of-the-week). Of these, one of the most sobering interviews is with glaciologist Eric Rignot of NASA, who speaks about his research on the melting Antarctic ice sheets (http://climatecrocks.com/2015/01/22/the-fuse-is-blown-glaciologists-jaw -dropping-account-of-a-shattering-moment/; 4:26 min). A second short video on the same page, made in conjunction with Yale Climate Connections, Meltwater Pulse 2B, models the melting of the Antarctic and Greenland ice sheets (5:56 min).

Climate change is arguably the most pressing existential issue for modern human civilization. An understandable response, given the scale of the threat, is to file climate change alongside other doomsday scenarios—like a gigantic meteor colliding with Earth—and hope it will not happen while privately marking it as unlikely. But unlike doomsday scenarios from science fiction, climate change is a challenge firmly grounded in scientific evidence. We can do something about it if we understand the problem clearly. However, more
research is needed on methods to successfully overturn people’s misconceptions about climate change: that is, how do we divorce the factual nature of this information from a sense of threatened cultural or political identity.

ACKNOWLEDGMENTS
We thank A. Malcolm Campbell for critical comments and helpful suggestions on this Feature and John Cook for useful discussion about his research on misconceptions.

AUTHORS’ NOTE
The number of results retrieved from querying databases was accurate as of March 2016, when the manuscript was written.

EDITOR’S NOTE
Based on the Plutzer et al. (2016a,b) report showing that many of today’s teachers are not educating America’s children about what is known about climate change, please consider sharing this Feature with friends and family members in educational positions.

REFERENCES

**Feature**

*From the National Academies*

**Insights from a Convocation: Integrating Discovery-Based Research into the Undergraduate Curriculum**

Sarah C. R. Elgin,† Gita Bangera,‡ Sean M. Decatur,§ Erin L. Dolan,‖ Laura Guertin,¶ Wendy C. Newstetter,∥ Elvyra F. San Juan,¶ Mary A. Smith,** Gabriela C. Weaver,†† Susan R. Wessler,∥∥ Kerry A. Brenner,§§ and Jay B. Labov∥∥∥

† Department of Biology, Washington University in St. Louis, St. Louis, MO 63130; ‡ RISE Learning Institute, Bellevue College, Bellevue, WA 98007; § Office of the President, Kenyon College, Gambier, OH 43022; ‖ Texas Institute for Discovery Education in Science, University of Texas, Austin, TX 78712; ¶ Department of Earth Science, Penn State Brandywine, Media, PA 19063; † College of Engineering, Georgia Institute of Technology, Atlanta, GA 30332; ‡ Office of the Chancellor, California State University System, Long Beach, CA 90802; ** Department of Biology, North Carolina A&T State University, Greensboro, NC 27411; †† Department of Chemistry, University of Massachusetts, Amherst, MA 01003; ‡‡ Department of Botany and Plant Sciences, University of California, Riverside, Riverside, CA 92521; §§ Board on Science Education and ††† Board on Life Sciences, National Academies of Sciences, Engineering, and Medicine, Washington, DC 20001

**INTRODUCTION**

Recommendation 2 of the February 2012 report *Engage to Excel* from the President’s Council of Advisors on Science and Technology (PCAST, 2012) urges the science, technology, engineering, and mathematics (STEM) education community and funding agencies to “advocate and provide support for replacing standard laboratory courses with discovery-based research courses.” The report justifies this recommendation as follows:

Traditional introductory laboratory courses at the undergraduate level generally do not capture the creativity of STEM disciplines. They often involve repeating classical experiments to reproduce known results, rather than engaging students in experiments with the possibility of true discovery.... Engineering curricula in the first two years have long made use of design courses that engage student creativity. Recently, research courses in STEM subjects have been implemented at diverse institutions, including universities with large introductory course enrollments. These courses make individual ownership of projects and discovery feasible in a classroom setting, engaging students in authentic STEM experiences and enhancing learning and, therefore, they provide models for what should be more widely implemented. (pp. iv–v)

This recommendation has engendered wide discussion and motivated formation of a committee at the National Academies of Sciences, Engineering, and Medicine to organize a convocation to explore opportunities and challenges of developing, implementing, and sustaining course-based undergraduate research experiences (CUREs), one mechanism for reaching large numbers of undergraduates. This column reviews the considerations leading up to the May 2015 convocation and summarizes the report that emerged (National Academies of Sciences, Engineering, and Medicine, 2015).1

Emerging evidence (cited in the PCAST report and elsewhere) indicates that engaging students in research as early as possible during their undergraduate years is one of the best strategies for supporting and retaining undergraduates in STEM. Until recently, undergraduates have primarily participated in research through apprenticeships, wherein an individual faculty member (or one of their graduate students or postdoctoral fellows) supervises the work of one or several students. Apprenticeships can be beneficial and even life and career changing for many students, yet their one-on-one design inherently limits the number of students who can participate. Providing all beginning STEM students with an individualized mentored
research experience in an apprentice-style program is not possible, given the large numbers of students. Faculty members (and other potential mentors from industry or elsewhere) have limited time, space, and resources to support undergraduates in their research activities. Most institutions have allocated only enough human and financial resources to involve a small fraction of their undergraduates in such experiences (PCAST, 2012). Students who seek out such positions are generally those who are already interested in research. Competition for a limited number of slots excludes many students, including students who have little knowledge of science career structures or who may not have performed well in traditional academic studies but who are capable of engaging in a research experience. For all of these reasons, students from populations historically underrepresented in STEM fields may especially be precluded from gaining an apprentice-style research experience (e.g., National Research Council 2007, 2011; Locks and Gregerman, 2008; National Academies of Sciences, Engineering, and Medicine, 2016). The PCAST report urges that access to a research experience during the undergraduate years should be seen as a pedagogical necessity for all students rather than a privilege for a small number of undergraduates.

Acting on this recommendation, many undergraduate STEM educators have been experimenting with different strategies for engaging more students in research experiences. There are now a number of different tested and emerging models available, and their successes suggest that this broad goal can be achieved through CUREs. Traditional “cookbook” laboratories are being replaced with discovery-based research and related activities. These are occurring both in laboratories associated with lecture courses and in stand-alone laboratory courses, using on-campus, off-campus, and online resources. A CURE allows undergraduates to engage in research either collectively or individually as part of a regularly scheduled course. Recent efforts have been aided by the Web, which can provide access to large data sets in topics from genomics to environmental monitoring, remote access to research-grade instruments, access to the scientific literature, and the means to build project consortia and to link student and faculty researchers at dispersed sites.

If appropriately constructed, managed, and mentored, CURES may be able to provide students with many of the same benefits acquired from summer apprentice-style research experiences (e.g., Shaffer et al., 2014). But CUREs also may provide additional advantages:

- By exposing more students to research and the nature of science earlier in their academic career, they can encourage students to explore the nature of various STEM topics and careers not otherwise considered.
- Course-based approaches may be a more effective and accessible starting point for many students, including minority, low-income, and first-generation college students (Bangera and Brownell, 2014).
- Faculty members may be able to undertake research that would otherwise be difficult or impossible (e.g., Leung et al., 2015; Pope et al., 2015).
- CUREs can provide all participating students with a greater ability to use scientific thinking in other aspects of their lives.

However, large-scale undergraduate research, especially for first- or second-year students, is not yet the norm on most campuses, particularly at those institutions with more traditional views of classroom teaching and reward and incentive systems or those with little or no on-site research. To make this recommendation from PCAST a reality, faculty and administrators need to be convinced of the feasibility, efficacy, cost efficiency, student benefits, and overall value of the approach. Because so many questions remain to be addressed regarding CUREs, the Board on Life Sciences, in collaboration with the Board on Science Education of the National Academies of Sciences, Engineering, and Medicine received support from three private foundations (see Acknowledgments) to develop an initiative that could further explore and elucidate the opportunities, barriers, and realities of CUREs as a potentially integral component of undergraduate STEM education. In response, a two-day national convocation was organized by a committee appointed by the National Academies and was held in Washington, DC, from May 11 to 13, 2015. Participants explored the following questions:

- What models have been developed to engage larger numbers of undergraduates in research using an academic year course-based format? Is this general strategy viable for all STEM disciplines and all class levels, from freshman to senior? Are minority-serving institutions participating, and are these models effective in reaching underrepresented students?
- Is the evidence base currently robust enough to identify best practices for implementation, considering different goals and different approaches? What are the most important challenges?
- Can these best practices serve as drivers of institutional cultural change, tackling some of the present barriers to access, and are there examples where they have done so?
- Is it possible to scale up to all students, without losing essential elements of the research experience?
- How do we promote and insure access and equity for all students in such initiatives?
- Can we recommend best practices for dissemination, for “start-up” support? What are the most cost-effective strategies?
- Can a shared research agenda help resolve some of these questions?

A common theme—equity and access issues for all students, with an emphasis on students from those populations that historically have been underrepresented in STEM—was emphasized throughout the convocation. The presentations and discussions that occurred at this convocation resulted in a summary report (Figure 1). The organizing committee selected 12 existing CUREs, presented as case studies, and others were described during the panel discussions.
the three foundations that supported the convocation also have provided financial support for follow-up dialogue through workshops at meetings of various scientific societies, including the American Society for Cell Biology and the Genetics Society of America. The convocation, publication, and follow-up dialogue are all helping to inform a larger, more in-depth National Academies consensus study on undergraduate research experiences, including all mechanisms—CUREs, apprentice models, internships, and others. That study is being supported by the National Science Foundation’s Division of Undergraduate Education. The committee’s report should be available by Fall 2016.

OVerview of Convocation Themes

Multiple themes emerged during the convocation, including those laid out by the committee (see bullet points above) and others raised by other participants. An important topic throughout the convocation was the critical characteristics of CUREs. Dr. David Asai from the Howard Hughes Medical Institute proposed that students should know that they are engaged in working on a real scientific problem, that their work matters to the community, and how their discoveries are contributing to the field. According to Asai, CUREs should permit students to encounter and confront problems that are important and timely; their work should contribute to advancing or refining knowledge, rather than simply repeating or “rediscovering” something that is already known. How to structure CURES that can help students advance to this level of discovery, particularly students who enroll in CUREs early in their undergraduate careers and may have only a single exposure to this kind of experience, engendered a great deal of discussion.

In his address as the keynote speaker at the convocation, Dr. James Gates, professor of physics at the University of Maryland, College Park, a member of the National Academy of Sciences and cochair of the committee that authored Engage to Excel (PCAST, 2012), provided a historical perspective of the relationship of research in science and technology to the nation’s economy and well-being (chap. 2 of convocation report). During the latter half of the 19th century and most of the 20th century, average educational levels in the United States were higher than in other countries, which he argued fueled the economic engine of this country. Educational levels of people in the United States are now lower than those in most other developed countries, and median household income has fallen, especially during the past 30 yr. Given the increase in available information and changes in the nature of work, Gates argued that today’s workers will need to continually relearn and retool their skills over their working lifetimes. He emphasized and provided details about how emerging approaches to education, including efforts to allow larger numbers of students to engage in discovery-based research, can help students advance to this level of discovery, particularly students who enroll in CUREs early in their undergraduate careers and may have only a single exposure to this kind of experience, engendered a great deal of discussion.

In this Feature, the authors describe activities at the National Academies aimed at understanding models and mechanisms for engaging undergraduates in research at scale and identifying the current state of knowledge and practice related to undergraduate research. We hope that the convocation report will stimulate broader conversations about the role of undergraduate research experiences for all students in undergraduate STEM education. Accordingly, the list of meetings where these dialogues will occur as of the date of publication of this article and copies of presentations from sessions already completed are available at http://dels.nas.edu/global/bls/Year-of-Dialogue.

Additional information about this study is available at http://sites.nationalacademies.org/DBASSE/BOSE/CurrentProjects/DBASSE_090473.
<table>
<thead>
<tr>
<th>Initiative (page numbers in convocation report)</th>
<th>Discipline(s) targeted</th>
<th>Local or national in scope?</th>
<th>Brief comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Science Education Alliance—Phage Hunters (SEA-PHAGES) (p. 18)</td>
<td>Life sciences</td>
<td>National collaborative</td>
<td>First-year students isolate bacteriophages whose genomes are then sequenced and annotated. Students routinely discover new viruses. Data are collected from multiple sites and made available to all participants. <a href="http://seaphages.org">http://seaphages.org</a></td>
</tr>
<tr>
<td>Genome Consortium for Active Teaching (GCAT) (p. 23)</td>
<td>Synthetic biology</td>
<td>National, but with faculty focus on specific topics of interest to them</td>
<td>Data from individual campuses are shared across the initiative. <a href="http://www.bio.davidson.edu/113/113labschedule2015.html">www.bio.davidson.edu/113/113labschedule2015.html</a></td>
</tr>
<tr>
<td>Genomics Education Partnership (pp. 24–25)</td>
<td>Genomics, bioinformatics, and evolution</td>
<td>National collaborative</td>
<td>Focuses on research around <em>Drosophila</em> genomes, including sequence improvement, annotation, and analysis of regions of interest. <a href="http://gep.wustl.edu">http://gep.wustl.edu</a></td>
</tr>
<tr>
<td>Genome Solver (p. 25)</td>
<td>Microbial genomics and bioinformatics</td>
<td>National collaborative</td>
<td>Focuses on analyzing data from the NIH Microbiome initiative. <a href="http://www.genomesolver.org">www.genomesolver.org</a></td>
</tr>
<tr>
<td>Place-Based Research (pp. 32–33)</td>
<td>Research on campus-based issues</td>
<td>Currently courses at the University of Wisconsin–Madison</td>
<td>Students conduct research on university issues such as energy use, food supply chains, and waste streams in collaboration with the University’s Office of Sustainability. <a href="http://nelson.wisc.edu/undergraduate/sustainabilitycertificate/syllabi/env_st_126-spring_2015_syllabus.pdf">http://nelson.wisc.edu/undergraduate/sustainabilitycertificate/syllabi/env_st_126-spring_2015_syllabus.pdf</a></td>
</tr>
<tr>
<td>Expanding the Use of Online Remote Electron Microscopy in the Classroom to Transform Undergraduate Education (pp. 34–36)</td>
<td>Geosciences</td>
<td>Currently a partnership among four colleges and universities in Florida</td>
<td>Students send samples to instrumentation labs and conduct individual research projects using the data obtained. <a href="http://fcaem.fiu.edu/tues">http://fcaem.fiu.edu/tues</a></td>
</tr>
<tr>
<td>DNA Learning Center (pp. 36–40)</td>
<td>DNA barcoding</td>
<td>Cold Spring Harbor Laboratory, New York</td>
<td>Students employ DNA barcoding to examine and resolve many kinds of questions, including identification of species in a habitat, food labeling. <a href="http://www.dnalc.org">www.dnalc.org</a></td>
</tr>
<tr>
<td>Science Education for New Civic Engagements and Responsibilities (SENCER) (p. 39)</td>
<td>Disciplinary, multidisciplinary, and interdisciplinary</td>
<td>National, with regional hubs that focus on different research issues</td>
<td>Students engage in research based on regional or global issues whose boundaries often extend beyond the STEM disciplines. <a href="http://sencer.net">http://sencer.net</a>; <a href="http://www.sencer.net/Resources/models.cfm">www.sencer.net/Resources/models.cfm</a></td>
</tr>
<tr>
<td>Virtual Internships (pp. 40–42)</td>
<td>Multidisciplinary and interdisciplinary</td>
<td>Individualized by instructors</td>
<td>Provides simulated experiences that give students the opportunity to both take and reflect on their actions to solve real-world problems, and develop ways of thinking about real-world practice. <a href="http://news.wisc.edu/uw-to-offer-new-virtual-internships-to-enhance-womens-interest-in-engineering">http://news.wisc.edu/uw-to-offer-new-virtual-internships-to-enhance-womens-interest-in-engineering</a></td>
</tr>
<tr>
<td>Vertically Integrated Projects (pp. 41–42)</td>
<td>Engineering, multidisciplinary and interdisciplinary</td>
<td>Currently 15 colleges and universities</td>
<td>Undergraduates have opportunities to participate in research for up to 3 years and help mentor newer members of the group. <a href="http://www.vip.gatech.edu">www.vip.gatech.edu</a></td>
</tr>
<tr>
<td>Rock Art Sustainability Index (RASI) (pp. 46–47)</td>
<td>Anthropology</td>
<td>Mesa Community College, Arizona</td>
<td>Students use the RASI to help determine which local Native American rock panel artworks are in greatest danger of eroding to allow for decisions about how to best use resources for preservation. In collaboration with the National Park Service. <a href="http://alliance.la.asu.edu/rockart/stabilityindex/RASI_Overview.html">http://alliance.la.asu.edu/rockart/stabilityindex/RASI_Overview.html</a></td>
</tr>
<tr>
<td>Center for Authentic Science Practice in Education (CASPIE) (pp. 48–51)</td>
<td>Chemistry</td>
<td>Currently 5 centers</td>
<td>Offers guidance on developing, implementing, and evaluating course-based models of research for first- and second-year undergraduates in chemistry. <a href="http://www.purdue.edu/discoverypark/caspie">www.purdue.edu/discoverypark/caspie</a></td>
</tr>
<tr>
<td>Scaffolding for Undergraduate Biology Using Yeast (p. 49)</td>
<td>Biological sciences</td>
<td>North Carolina Central University</td>
<td>Designed to involve students underrepresented in STEM fields with authentic research experiences as part of three introductory biology courses. <a href="http://www.phdadavid.com/documents/McDonaldCUREnet2014.pdf">www.phdadavid.com/documents/McDonaldCUREnet2014.pdf</a></td>
</tr>
<tr>
<td>The Nature of Life (pp. 51–54)</td>
<td>Life sciences</td>
<td>University of Minnesota–Twin Cities</td>
<td>A required 2-yr, two-credit course for all entering biology majors that starts in the summer before the students’ first year at the university. <a href="http://cbs.umn.edu/academics/departments/btl/academics/nol-series">http://cbs.umn.edu/academics/departments/btl/academics/nol-series</a></td>
</tr>
</tbody>
</table>

(Continued)
research, can help meet the expanding need for workers trained in STEM fields. These points served as the basis for recommendation 2 in the PCAST report.

**Assessment and Evaluation of CUREs**

A plenary session focused on what emerging research indicates about the efficacy of CUREs on several levels. Student and faculty enthusiasm for CUREs is, at present, largely based on student reports of learning gains and satisfaction with the experience (Auchincloss et al., 2014; Corwin et al., 2015; Linn et al., 2015). However, there are some well-documented studies showing that research experiences improve retention in the sciences (e.g., Locks and Gregerman, 2008; Estrada et al., 2011; Schultz et al., 2011; Eagan et al., 2013; summarized in Corwin et al., 2015), and several case studies presented at the convocation reported positive impacts. CURE assessments that use multiple indicators of student learning and program efficacy can provide greater insights concerning achievement of desired learning goals and affective behaviors of students and can offer guidance when starting new courses (Corwin et al., 2015; Linn et al., 2015; more research of this type is needed.

Many CUREs are designed by individual faculty to align with their own research interests, an approach that has many benefits but results in assessments that are idiosyncratic and difficult to compare (Lopatto, 2010; Linn et al., 2015). In contrast, a group of coordinated national efforts (Table 1) have attempted to address these issues by using common assessments, and some positive results have been reported (Jordan et al., 2014; Shaffer et al., 2014). Speakers pointed out that collaborative projects and/or cooperatives of schools with common program goals and common sets of activities can develop a common set of metrics, providing unique opportunities for assessing their efforts. Moreover, speakers noted the potential for partnerships among state systems of higher education and public and private consortia for fostering the acceptance and institutionalization of research-based courses.

### Table 1. Continued

<table>
<thead>
<tr>
<th>Initiative (page numbers in convocation report)</th>
<th>Discipline(s) targeted</th>
<th>Local or national in scope?</th>
<th>Brief comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freshman Research Initiative (pp. 52–53)</td>
<td>Multiple disciplines</td>
<td>University of Texas–Austin</td>
<td>Offers first-year students in the College of Natural Sciences an opportunity to conduct original research under the guidance of a research faculty member and graduate students through a three-semester sequence of courses and laboratory work. <a href="https://cns.utexas.edu/fri">https://cns.utexas.edu/fri</a></td>
</tr>
<tr>
<td>Community College Undergraduate Research Initiative (pp. 56–57)</td>
<td>Multiple disciplines</td>
<td>National</td>
<td>Exposes community college students to real-world science through hands-on research experiences. Students take an introductory course in which they are taught basic scientific procedures while investigating a specific case study and then work together to investigate questions developed from a case study. <a href="http://www.ccuri.org/content/home">www.ccuri.org/content/home</a></td>
</tr>
<tr>
<td>Discovery-Enriched Curriculum (pp. 61–63)</td>
<td>All disciplines</td>
<td>City University of Hong Kong</td>
<td>Institution-wide program that requires all 11,000 students who matriculate to make an original discovery or create intellectual property. <a href="http://www.cityu.edu.hk/provost/dec">www.cityu.edu.hk/provost/dec</a></td>
</tr>
<tr>
<td>Interdisciplinary Science Learning Labs (pp. 63–65)</td>
<td>All disciplines</td>
<td>University of Delaware</td>
<td>Engages undergraduates in all phases of scientific research through the development of facilities that foster the integration of teaching, learning, and research in a holistic learning environment. <a href="http://www.udel.edu/iselab">www.udel.edu/iselab</a></td>
</tr>
<tr>
<td>Center for Interdisciplinary Biological Inspiration in Education and Research (CIBER) (p. 64)</td>
<td>Engineering design inspired by biological structures and functions</td>
<td>University of California–Berkeley</td>
<td>Creates a community of next-generation scientists and engineers who can work together to conceive and execute innovative multidisciplinary work by engaging undergraduates to formulate and execute novel designs in engineering that are informed and inspired by biological principles and phenomena. <a href="http://ciber.berkeley.edu">http://ciber.berkeley.edu</a></td>
</tr>
<tr>
<td>First-Year Innovation and Research Experience (FIRE) (pp. 65–68)</td>
<td>All disciplines</td>
<td>University of Maryland–College Park</td>
<td>Modeled after the Freshman Research Initiative at the University of Texas (see description above), FIRE provides first-year students with authentic research experiences, broad mentorship, and institutional connections, but with an expansion to disciplines beyond the STEM fields. <a href="http://fire.umd.edu">http://fire.umd.edu</a></td>
</tr>
<tr>
<td>Dynamic Genome Project (pp. 66–67)</td>
<td>Genomics and molecular biology</td>
<td>University of California–Riverside</td>
<td>Provides undergraduates with the same types of experimental activities as graduate students while they learn fundamental concepts in genomics and molecular biology in a two-course sequence that is required for biology majors. <a href="http://dynamicgenome.ucr.edu">http://dynamicgenome.ucr.edu</a></td>
</tr>
</tbody>
</table>

---

*Given the limited amount of time to address many topics during the convocation, no topic was explored in detail. The consensus study now underway at the National Academies of Sciences, Engineering, and Medicine will address many of these issues more deeply. A primary charge to that committee is to examine the robustness of the research literature on assessment of CUREs and other types of undergraduate research experiences.*
Leveraging Resources to Provide More Students with Opportunities for Research

A great deal of discussion at the convocation centered on issues of resources and costs of CUREs compared with more traditional teaching laboratories or apprentice-based models of research participation. There are surprisingly few data available on relative costs, and where available, most of the information is incomplete; for example, faculty are often not compensated for mentoring students under the apprenticeship model, so this cost is not recorded, but will be compensated for teaching a CURE. Benefits are also difficult to monetize. However, if CUREs can be demonstrated to increase retention of students, either in STEM or more broadly at the institution, the money from those students’ tuition can financially justify the strategy. Several presenters noted that research-based courses were powerful recruiting and retention tools for the departments or colleges that offer them. Several presenters reported that costs for running CUREs, even if slightly more expensive than traditional labs, could be borne by students through a small increase in lab fees. However, undergraduates who participated in the convocation pointed out that even a small increase in fees can be a much larger amount if included in high-interest student loans and could make a difference in their decision to enroll in a discovery-based lab course versus a more traditional one. Some people claimed that the CUREs they designed were less expensive; techniques such as DNA barcoding and synthetic biology open up many avenues at modest cost.

Several presenters emphasized that costs for CUREs can be reduced by taking advantage of local resources and partnerships. Specialized analytical instruments can be accessed remotely, often at low cost. Further, every campus needs to collect and analyze data on the operation of their physical plant, on the use of services, and so on, and students can be engaged in the effort. For example, California State University’s recent initiative, “The Campus as a Living Laboratory,” and student research undertaken with the University of Wisconsin’s food services to better understand food use and waste by students were both described. Some of the costs recovered can be directed back into the student research program (Cathy Middlecamp, University of Wisconsin–Madison, personal observation).

Opportunities and Challenges to Scaling CUREs

The convocation provided many examples of how to restructure a given course as a CURE, but we have little experience in “scaling up” these efforts. Expanding such efforts to all sections of a course or most courses within a department may seem desirable, but the logistics and infrastructure required to do so may seem prohibitive. A great deal of discussion during the convocation thus focused on issues of scaling up of CUREs. Because appropriate mentoring is hard to provide for large numbers, it was pointed out by David Shaffer (University of Wisconsin–Madison) that virtual internships (online challenges that prompt students to take action on a complex problem, reflect on their action, and develop ways of thinking about real-world practice) can enable many more students to participate, reaching more students than would be possible through internships in which students must be at the site of the internship to participate. However, several universities (e.g., University of Minnesota–Twin Cities, University of Texas–Austin; Table 1) are expanding wet-bench CUREs by organizing multiple research “streams,” among which a student can choose, using a hierarchical mentoring system. A parallel in engineering is the Vertically Integrated Projects program, described by Edward Coyle (Georgia Tech), in which senior students contribute to mentoring beginning students (Table 1).

Several convocation speakers indicated that research-based courses can help a broad range of students decide whether they would like to pursue additional research opportunities, arguing that at least one such course should be mandatory for all students. Mandatory participation ensures that students may lack the confidence to pursue such pathways on their own are able to do so (Bangera and Brownell, 2014). However, requiring a research experience also means that some students may feel that they are being forced into something they do not want; the undergraduates present at the convocation noted that students who are working hard to maintain or increase their grade point average may be wary of a process in which the probability of failure may be high and course grades are not based on the usual criteria. Hence, for some students, an effort must be made to show them how a research experience will benefit them.

In a panel on institutional strategies, Goldie Byrd (North Carolina A&T State University) pointed out that department chairs, deans, and other administrators can support CUREs by actively promoting faculty professional development in teaching and mentoring and by supporting faculty time used to develop a CURE. The construction of new instructional spaces or the reconfiguration of existing spaces for CUREs also offers opportunities to change the culture of teaching and learning, as seen in the new Interdisciplinary Science Learning Laboratories at the University of Delaware and described by John Jungck. Efforts toward the establishment of endowments and special funds can send powerful signals to faculty, students, regents, and parents about the value of this kind of work. For those campuses undertaking or contemplating major curricular reforms, open consideration of investing in CUREs may provide opportunities to rethink the integration of research into undergraduate education and to retool the reward system for faculty, a major change strategy utilized by the City University of Hong Kong and described by Arthur Ellis.

Closing remarks stressed that the creation of new knowledge is a major function of universities. By welcoming students into this effort, we make them our partners and provide a sense of belonging in this field. As Robin Wright said in describing the freshman program at the University of Minnesota, “We’re talking to them as if they are emerging professional biologists, and we treat them as colleagues.” Teaching STEM by having students do research can be viewed as an active-learning strategy—and there is considerable evidence that active-learning strategies work (e.g., Freeman et al., 2014). The convocation closed with enthusiasm for using CUREs to expand research opportunities to all students. Jim Gentile, now at the University of Arizona, concluded that

---

7 Additional information available at www.calstate.edu/cpdc/sustainability/liv-lab-grant.
“undergraduate research is quality education.” Collectively, the speakers and discussions engendered enthusiasm for several elements related to CUREs, and we hope that this report will be useful for those faculty and schools thinking about or planning to expand the use of CUREs in their curriculum.

ACKNOWLEDGMENTS

The convocation and year of dialogue have been generously supported by the Alfred P. Sloan Foundation, the Howard Hughes Medical Institute, the Leona M. and Harry B. Helmsley Charitable Trust, and the President’s Committee of the National Academies of Sciences, Engineering, and Medicine.

REFERENCES


Essay

A Conceptual Framework for Graduate Teaching Assistant Professional Development Evaluation and Research

Todd D. Reeves,†* Gili Marbach-Ad,‡ Kristen R. Miller,§ Judith Ridgway,‖ Grant E. Gardner,¶ Elisabeth E. Schussler,# and E. William Wischusen@

†Educational Research and Assessment, Northern Illinois University, DeKalb, IL 60115; ‡College of Computer, Mathematical and Natural Sciences, University of Maryland, College Park, MD 20742; §Division of Biological Sciences, University of Georgia, Athens, GA 30602; ‖Center for Life Sciences Education, Ohio State University, Columbus, OH 43210; ¶Department of Biology, Middle Tennessee State University, Murfreesboro, TN 37132; #Ecology & Evolutionary Biology, University of Tennessee, Knoxville, TN 37996; @Biological Sciences, Louisiana State University, Baton Rouge, LA 70803

Submitted October 28, 2015; Revised March 16, 2016; Accepted March 19, 2016
Monitoring Editor: Marilyne Stains

Biology graduate teaching assistants (GTAs) are significant contributors to the educational mission of universities, particularly in introductory courses, yet there is a lack of empirical data on how to best prepare them for their teaching roles. This essay proposes a conceptual framework for biology GTA teaching professional development (TPD) program evaluation and research with three overarching variable categories for consideration: outcome variables, contextual variables, and moderating variables. The framework’s outcome variables go beyond GTA satisfaction and instead position GTA cognition, GTA teaching practice, and undergraduate learning outcomes as the foci of GTA TPD evaluation and research. For each GTA TPD outcome variable, key evaluation questions and example assessment instruments are introduced to demonstrate how the framework can be used to guide GTA TPD evaluation and research plans. A common conceptual framework is also essential to coordinating the collection and synthesis of empirical data on GTA TPD nationally. Thus, the proposed conceptual framework serves as both a guide for conducting GTA TPD evaluation at single institutions and as a means to coordinate research across institutions at a national level.

Biology graduate teaching assistants (GTAs) have important instructional roles in undergraduate education at colleges and universities. Rushin et al. (1997) reported that of 153 surveyed graduate schools, 97% used GTAs in some form of undergraduate instructional role. In another study, Sundberg et al. (2005) reported that biology GTAs teach 71% of laboratory courses at comprehensive institutions and 91% of laboratory courses at research institutions. More recently, a national survey of 85 faculty and staff providing teaching professional development (TPD) to biology GTAs found that 88% of those surveyed were preparing GTAs to teach introductory-level biology courses (Schussler et al., 2015). Thus, GTAs have a potentially powerful impact on undergraduate student learning at many colleges and universities, especially in introductory laboratories and introductory-level lecture courses.

Introductory science courses are often the “gateway” to the attainment of undergraduate science degrees, and progression through the degree and beyond often depends on undergraduate student performance in these early courses (Seymour and Hewitt, 1997). This makes these courses uniquely important for student retention as the nation attempts to increase the number of science, technology, engineering, and mathematics (STEM) graduates (President’s Council of Advisors on Science and Technology, 2012). Biology education researchers have argued that because of the smaller and more intimate class size of introductory-course laboratory and discussion sections, GTAs contribute meaningfully to retention efforts, because they have
more personal contact with first-year students than do most faculty members (Rushin et al., 1997). Providing biology GTAs with opportunities to develop instructional expertise that maximizes student learning outcomes should be a priority for the universities that employ them, yet GTA teaching responsibilities are often relegated to secondary status or sometimes even actively discouraged (Nyquist et al., 1999; Gardner and Jones, 2011).

Currently, there is wide variation among universities and departments vis-à-vis biology GTA TPD. A recent national survey found that 96% of responding TPD practitioners provided some formal TPD to their biology GTAs (e.g., TPD workshop) but that these programs varied extensively in terms of total contact hours (2–100 h per academic year). Because many of these contact hours are delivered as one-time presemester workshops between 2 and 5 h in length (Schussler et al., 2015), GTA TPD does not generally meet research-based TPD standards (Garet et al., 2001; Desimone et al., 2002). Institutional differences in the levels of funding and support for TPD programs (Schussler et al., 2015) suggest that university and/or department contextual variables may impact TPD design quality (as suggested by Park, 2004; Seymour et al., 2005).

The current state of biology GTA TPD highlights the need for further research on biology GTA TPD that accounts for the diverse institutional contexts in which these TPD programs are implemented. This mirrors recent calls for “biology education research 2.0” to better consider contextual factors (Dolan, 2016). The current literature base for GTA TPD is primarily limited to small-scale evaluation studies concerning individual TPD programs (Abbott et al., 1989; Marbach-Ad et al., 2015a). Though these studies can be used to suggest practices that TPD leaders may adopt, there is no guarantee that what worked at one institution will effectively transfer to a different context. At the same time, existing studies often do not compare the efficacy of different TPD practices and frequently use different assessment tools, making cross-institutional and cross-study comparisons difficult. A systemic approach to evaluation and research is needed to identify evidence-based practices in biology GTA TPD.

This article proposes a conceptual framework for GTA TPD evaluation and research suggesting that the most important TPD program outcomes to measure (as determined by our BioTAP1 working group and the current literature) are GTA cognition, GTA teaching practice, and undergraduate student outcomes. The framework also highlights key contextual variables that should be considered in broad-scale examinations of GTA TPD and potential moderators of TPD impact. It builds on the model put forth by DeChenne et al. (2015) but is more global in nature, positing the importance of multiple categories of relevant GTA TPD variables. The intent of this framework, then, is to support TPD practitioners in the evaluation of their programs (on their own or with assistance from an educational researcher/evaluator). At the same time, the framework provides a structure for cross-institutional collaborations focused on the conduct, synthesis, and dissemination of research related to evidence-based biology GTA TPD practices. Crucially, this essay also offers categories of instrumentation and examples of specific instruments that GTA TPD practitioners might use in local and large-scale GTA TPD evaluation and research.

### EVALUATION OF GTA TPD PROGRAMS

Given long-standing concerns that GTA TPD is inadequate (Boyer Commission on Undergraduates in the Research University, 1998; Gardner and Jones, 2011), evaluation of GTA TPD programs is critical. Such efforts can ensure TPD program effectiveness and/or the refinement of programs to support and enhance the quality of GTA teaching and, as a result, the learning outcomes of undergraduates. When discussing evaluation, the literature recognizes two overarching types of evaluations that are differentiated by their purpose: formative and summative (Patton, 2008; Yarbrough et al., 2010).

In this context, at its core, formative evaluation endeavors to inform iteratively the quality of GTA TPD program design and implementation. As an example of a formative evaluative activity, a GTA TPD program staff member might collect data after the first of two TPD sessions to identify content GTAs would like to revisit during the third TPD session (Marbach-Ad et al., 2015a). Summative evaluation, on the other hand, aims to summarize what happened as a result of GTA TPD program implementation. For example, researchers might seek to describe whether a GTA TPD program was associated with increased inquiry-based teaching in laboratories (e.g., Ryker and McConnell, 2014). It is also noteworthy that a particular GTA TPD program evaluation effort can serve both formative and summative purposes. For example, an end-of-TPD summative evaluation can inform the design of the next semester’s TPD program. Many of the key constructs in the conceptual framework proposed herein (e.g., GTA cognition, GTA teaching practice) can be examined for formative purposes, summative purposes, or both.

An expedient means of formatively evaluating a GTA TPD program is through the collection of data concerning GTA participants’ satisfaction. Measures of satisfaction capture how the respondent feels or thinks about the program. For example, an evaluation might ask GTAs who participated in a TPD program the degree to which they were satisfied with the program as a whole and/or with its particular components, activities, or processes (e.g., lectures, group activities, microteaching). Satisfaction is also commonly assessed at the end of GTA TPD programs, typically via post-TPD surveys, for summative evaluation purposes (e.g., Baumgartner, 2007; Vergara et al., 2014). However, researchers have long criticized GTA satisfaction as an appropriate measure of outcomes in GTA TPD intervention research (Chism, 1998; Seymour, 2005), because the relationship between participants’ satisfaction and actual learning is equivocal at best (e.g., Gessler, 2009). Therefore, while we recognize the use of satisfaction in the GTA TPD literature, we do not include it in our evaluation and research framework, because we argue it is a fundamentally different variable than program outcomes such as GTA cognition, GTA teaching practice, and undergraduate student outcomes.

---

1The Biology Teaching Assistant Project (BioTAP) and the Biology Teaching Assistant Project: Advancing Research, Synthesizing Evidence (BioTAP 2.0) are, respectively, a National Science Foundation–funded Research Coordination NETWORK Incubator (DBI-1247938) and a National Science Foundation–funded Research Coordination Network (DBI-1539903).
Graduate Teaching Assistants

CONCEPTUAL FRAMEWORK

Figure 1 presents our proposed conceptual framework for evaluation and research related to GTA TPD programs. The purpose of this framework is twofold: 1) to guide those who are planning to conduct empirical evaluation or research studies related to a particular GTA TPD program (at a particular department, college, or institution); and 2) to guide researchers interested in conducting, synthesizing, and disseminating large-scale and multisite research on GTA TPD.

The framework hypothesizes several categories of variables that are related to the operation of GTA TPD and is based on extant theory and research on GTA TPD (e.g., DeChenne et al., 2015) and on broader conceptual frameworks for evaluation of professional development programs (e.g., Guskey, 2000; Wyse et al., 2014). The framework contains three categories of variables: outcome variables, contextual variables, and moderating variables. In Figure 1, we provide nonexhaustive examples of key variables in each of these categories.

OUTCOME VARIABLES

An essential focus of GTA TPD program evaluation and research is on a program’s outcomes relative to its goals and objectives. The proposed framework contains three
main categories of outcomes (or impacts) that programs may measure (blue in Figure 1): GTA cognition, GTA teaching practice, and undergraduate student outcomes. Two of these outcomes pertain to GTAs and one outcome pertains to undergraduate students. Moreover, these outcome variable categories are linearly (sequentially) related, in that TPD directly impacts GTA cognition, which in turn impacts GTA teaching practice, which then impacts undergraduate student outcomes.

GTA Cognition

GTA cognition pertains to cognitive changes in GTAs’ knowledge, skills, and attitudes toward or beliefs about teaching that directly result from the GTA TPD. For example, such outcomes might include GTA knowledge of active learning or inquiry-based teaching techniques or GTA teaching self-efficacy beliefs (e.g., Bowman, 2013; Connolly et al., 2014). Hardré (2003) and DeChenne et al. (2015) reported evidence for a relationship between participation in TPD and GTA cognition (i.e., knowledge and self-efficacy).

GTA Teaching Practice

GTA cognition is linked to GTA teaching practice, which concerns GTAs’ behavior related to planning, instruction, and assessment. Prior research, for example, documented improvements in GTA instructional planning and assessment practices as a result of TPD (Baumgartner, 2007; Marbach-Ad et al., 2012), and Hardré (2003) linked GTA cognition (self-efficacy) and instructional practice in the context of GTA TPD. Generally, examination of GTA teaching practices will focus on teaching practices that were discussed in the GTA TPD. For example, if one of the TPD goals is to enhance inquiry-based teaching in the laboratory, part of the evaluation/research activities will focus on the level and adequacy of the implementation of inquiry-based instruction.

Undergraduate Student Outcomes

Finally, undergraduate student outcomes center on the gains in knowledge and skills made by GTAs’ students, as well as more distal student outcomes such as retention and graduation. For example, one might expect that undergraduate students taught by GTAs who have received TPD would perform better on course exams. Indeed, research in K–12 settings has found that measures of teacher self-efficacy (a cognitive belief) are related to both teaching practices and student achievement (Tschanne-Moran et al., 1998).

In sum, the framework uses existing literature to posit that GTA TPD directly promotes changes in participants (GTA cognition), which in turn affects their instructional behavior (GTA teaching practice), and, subsequently, outcomes for undergraduates (undergraduate student outcomes). Of these three GTA program outcomes, the first (GTA cognition) has been examined most often in GTA evaluation and research (unpublished data). Examination of the other two outcomes, GTA teaching practices and undergraduate student outcomes, is logistically more challenging and expensive, depending on the instrumentation used.

Multisite evaluation of these latter outcomes is furthermore challenging, owing to varying contextual factors (e.g., the roles of the GTAs, undergraduate course content). However, we contend that the most comprehensive and scientifically rigorous GTA TPD evaluation should consider each of these three outcomes (and employ true experimental or quasi-experimental designs in order to confidently assess whether changes in these variables are due to GTA TPD rather than other variables). For those just starting evaluations of their programs, it would be reasonable to start with the most proximal GTA TPD outcome (i.e., GTA cognition), and once those effects are established, proceed to the evaluation of more distal outcomes (i.e., GTA teaching practice, then undergraduate student outcomes). In a later section, we offer practical guidance on how to elicit evidence of various GTA TPD outcomes.

CONTEXTUAL VARIABLES

As mentioned earlier, one limitation of the GTA TPD literature is that it largely comprises small-scale studies, each focused on a particular GTA TPD program at a particular institution. As such, the literature lacks large-scale, multi-institutional studies with the potential to compare the effectiveness of GTA TPD programs that systematically vary in their design, allowing for identification of evidence-based practices (Hardré and Chen, 2005; Hardré and Burris, 2012). The challenge of drawing comparisons among different TPD designs from the extant literature is furthermore compounded by considerable variation among institutional contextual factors (Schussler et al., 2015). For example, findings from DeChenne et al.’s 2015 study underscored the importance of accounting for contextual variables, such as departmental teaching climate, when studying GTA TPD programs. Therefore, what might constitute an “effective” GTA TPD program for one institution/department might not be effective for another.

Given the generally fragmented nature of the body of GTA TPD literature, our framework considers three categories of contextual variables (in yellow in Figure 1): GTA training design variables, institutional variables, and GTA characteristic variables. These elements of the framework are intended for researchers interested in conducting research on GTA TPD program design and impact in diverse contexts. The categories also offer guidance for the types of information that individuals who publish outcomes of single GTA TPD programs should provide to situate the context of their program for their readers.

GTA Training Design Variables

The design of GTA TPD varies widely, in terms of training program content, structure, and activities (e.g., Hardré and Burris, 2012; DeChenne et al., 2015). In the proposed conceptual framework, GTA TPD training design variables are hypothesized to drive the most direct outcome of GTA TPD—GTA cognition. As noted earlier, GTA cognition ultimately affects GTA teaching practices and, in turn, undergraduate student outcomes. Notably, K–12 professional development designs that translate to teacher and/or student outcomes are marked by a focus on subject matter content, coherence with teachers needs (content), an extended duration (structure), and opportunities for active learning (activities; Garet et al., 2001; Desimone et al., 2002).

There is also some published literature on the design of GTA TPD in terms of its content, structure, and activities. With respect to TPD content, TPD programs described in the literature have covered topics such as assessment, pedagogical
methods, policies and procedures, and multicultural issues (e.g., Luft et al., 2004; Prieto et al., 2007). In terms of TPD structure, GTA TPD programs discussed in the literature often take the form of a onetime workshop (Gardner and Jones, 2011; Schussler et al., 2015); other designs or design elements such as GTA mentoring or receipt of teaching feedback are much more rare (Austin, 2002; DeChenne et al., 2012). Relative to TPD activities, prior research has examined activities such as microteaching (Gilreath and Slater, 1994) and teaching skits (Marbach-Ad et al., 2012). Published GTA TPD research even offers evidence for positive effects of some TPD design variables on GTA cognition, for example, the effect of training length on GTA self-efficacy related to teaching (e.g., Prieto and Meyers, 1999; Hardré, 2003; Young and Bippus, 2008).

**Institutional Variables**

The proposed conceptual framework incorporates institutional variables such as institutional type, size, student body characteristics, and policy training requirements. Institutional variables are hypothesized to have effects on the nature of the TPD provided to GTAs, although concrete empirical evidence for this is sparse and often indirect (Park, 2004; Lattuca et al., 2014). As noted previously in the literature, TPD content and structure vary considerably from institution to institution and across different institutional contexts (Marbach-Ad et al., 2015a; Schussler et al., 2015), including institutional cultural differences with respect to how teaching is viewed (Serow et al., 2002). Along these lines, Rushin et al. (1997) found differences between master’s degree- and doctoral degree-granting institutions in terms of the GTA TPD models used. In their study, doctoral degree-granting institutions were more likely to employ a preacademic-year workshop, whereas master’s degree-granting institutions were more likely to employ individualized GTA training led by the course professor. While the Rushin et al. (1997) findings are suggestive of a key role of institutional type (e.g., research-intensive university) in shaping GTA TPD design, (arguably) other variables are important as well. For example, the typical teaching role of the GTA at a particular institution (e.g., facilitating discussion sessions, coordinating laboratory sessions, or grading assignments) and the presence of a faculty development unit (e.g., Center for Teaching and Learning; Marbach-Ad et al., 2015a) might also affect the design of a GTA TPD program, specifically its duration, structure, or content.

**GTA Characteristic Variables**

Finally, a third category of contextual variable in the proposed framework is GTA characteristics. The extant literature highlights considerable variation among GTAs both across and within institutions (Addy and Blanchard, 2010; DeChenne et al., 2015). In particular, GTAs differ with respect to their prior teaching experiences and training (Prieto and Altmaier, 1994), relative prioritization of teaching versus research, aspirations for careers involving teaching (Nyquist et al., 1999; Brownell and Tanner, 2012; Sauermann and Roach, 2012), and attitudes toward teaching (Tanner and Allen, 2006). In the framework, GTA characteristics are posited to impact the nature of the TPD provided to GTAs (i.e., TPD training design). A GTA population with varying levels of teaching experience, for example, might necessitate a differentiated TPD program (Austin, 2002; Schussler et al., 2015). As another example, Marbach-Ad et al. (2015a,b) reported on three different TPD programs at their research-intensive university based on students’ career aspirations. Thus, GTA characteristics can impact GTA training design variables such as duration (e.g., a longer course for those with teaching aspirations), structure (e.g., type and amount of homework assignments), and activities (e.g., developing a teaching philosophy and portfolio).

GTA characteristics are also hypothesized to directly impact GTA cognition (e.g., knowledge/skills, attitudes, and beliefs) and GTA teaching practice, independent of TPD. Prior research indicates large GTA-to-GTA variation even after participation in TPD (e.g., Bond-Robinson and Rodrigues, 2006; Addy and Blanchard, 2010), implying that other GTA-level variables besides training (i.e., GTA characteristics) impact GTA teaching cognition and practice. For example, research has shown a relationship between GTA level of teaching experience and teaching self-efficacy (Prieto and Altmaier, 1994) and that diverse GTA beliefs and prior experience impact their teaching practices (Addy and Blanchard, 2010). Moreover, these GTA characteristics should be considered in the interpretation of GTA evaluation findings. For instance, when comparing the effectiveness of two programs, one needs to consider the GTAs’ input characteristics (e.g., prior TPD experience), because differential knowledge after training might be caused by those initial differences rather than differences in program effectiveness.

**MODERATING VARIABLES**

The proposed framework also includes two categories of moderator variables (in green in Figure 1): implementation variables and GTA characteristic variables. These variables are termed moderating variables, because they may impact or modify the relationship between two other variables (in this case the relationship between GTA training design and GTA cognition).

**Implementation Variables**

The success of any program in attaining its intended outcomes depends not only on the TPD program’s intended design but also on how well it was implemented. Evaluation of program implementation involves examining the degree to which a GTA TPD program was enacted with fidelity, that is, as intended. We therefore also included implementation variables (i.e., Dane and Schneider’s [1998] concepts of program adherence, exposure, and participant responsiveness) in the proposed framework as moderators of the relationship between TPD training design variables and GTA cognition outcomes. If null effects of GTA TPD are observed, implementation variable data (e.g., the number of times each GTA met with his or her mentor) can assist program staff in discerning whether effects were not observed because of a poorly designed program (i.e., theory failure) or poor program implementation (i.e., implementation failure). Examples of implementation variables that might be assessed include the GTAs’ degree of participation/engagement in the TPD program, the degree to which all intended content was given sufficient attention during a TPD session, or whether protocols for collaborative learning activities for GTAs were followed appropriately. This information is
often collected through the use of external observers during the program, but it could also be collected from GTAs’ self-reports during end-of-semester survey or interviews. For example, Marbach-Ad et al. (2015b) used an external evaluator to interview and survey GTAs who participated in a teaching certificate program. The design of the program included a component in which GTAs were observed and mentored by faculty members. GTAs reported that this component was not well implemented, mainly due to lack of faculty cooperation, suggesting that poor implementation might have moderated the relationship between GTA training design variables and TPD outcome variables.

**GTA Characteristic Variables**

The proposed framework also includes GTA characteristics as moderators of the relationship between GTA training design and GTA cognition. Simply put, this aspect of the framework pertains to possible **differential effects of TPD on GTA cognition**. Several studies have investigated the relationship between GTA prior teaching experience (e.g., number of semesters taught) and self-efficacy belief and attitudinal gains observed during TPD (e.g., Addy and Blanchard, 2010; DeChenne et al., 2015). Other work has shown that GTAs’ prior teaching experiences or knowledge is related to knowledge gains during TPD (Marbach-Ad et al., 2012) and to the implementation of TPD content during GTAs’ classroom practice (French and Russell, 2002; Hardré and Chen, 2005).

**APPLYING THE CONCEPTUAL FRAMEWORK**

Implicit in each of the proposed framework’s directional paths are various evaluation and research questions/hypotheses about how GTA TPD programs operate to produce GTA and student outcomes and about the role of contextual variables in GTA TPD. These include 1) system-level questions, such as how institutional variables affect GTA TPD training design; 2) TPD program-level questions, such as how different TPD training designs translate to direct effects on GTAs’ cognition and indirect effects on GTA teaching practices and undergraduate student outcomes; and 3) individual GTA-level questions, such as how GTAs with different characteristics respond differently to TPD. Through its inclusion of contextual variables, the framework also provides a structure for both small-scale, local (single program) evaluation and large-scale, cross-institutional GTA TPD research (looking across programs to identify evidence-based practices).

Even if a researcher is studying only a single, local GTA program and its outcomes, in reporting his or her findings, he or she should describe the program’s design, implementation, and relevant contextual variables in terms of the institution and participating GTAs. This will afford the community more information to use in synthesizing findings across individual studies. At the same time, such information can help a reader weigh the applicability of a given study’s findings to his or her local context. For example, findings derived from a TPD program for GTAs who want to enter industrial fields may not necessarily apply to a TPD program for GTAs who hope to attain positions at small, liberal arts colleges focused chiefly on teaching.

It bears noting that the framework is general in nature, in that it theorizes relationships between categories of variables (e.g., GTA training design and GTA cognition) rather than relationships between specific variables (e.g., GTA training length and GTA beliefs about teaching). Specific variables are provided for illustrative purposes. The framework does not posit that every specific variable represented within a particular variable category (a box in Figure 1) is associated with every specific variable represented within a related category. Continued research is needed to empirically elicit the relationships between specific variables in each general category.

While the proposed framework is inclusive of several key categories of variables, it is not exhaustive in the sense that all determinants of GTA TPD design, implementation, and outcomes are included. For instance, in addition to institutional and GTA characteristic variables, TPD program staff variables (e.g., knowledge, beliefs) might also impact GTA TPD design. As additional evidence accumulates, other welcomed extensions to the general framework described here may include mediators or moderators of particular linkages (e.g., student population moderating the impact of certain GTA classroom practices on student achievement, or GTA curricular autonomy moderating the impact of GTA cognition on GTA practice). We hope that future research validates this framework and refines it as needed on the basis of evidence.

**A PRACTICAL GUIDE FOR EVALUATING GTA TPD PROGRAMS**

In Table 1, we offer practical guidance for those who wish to conduct evaluations of their own GTA TPD programs. In particular, we discuss how to elicit evidence of the three GTA TPD outcome variables implicit in the proposed conceptual framework (GTA cognition, GTA teaching practice, and undergraduate student outcomes). For each of these three GTA TPD outcomes, we enumerate some guiding evaluation questions, possible categories of instrumentation (e.g., surveys, tests), and examples of specific existing instruments (e.g., Smith et al.’s [2008] Genetics Concept Assessment) that can be used in evaluation efforts. We caution that the specific instruments we reference are provided as examples but may not be the most appropriate for any given program.

In addition, we recommend that researchers interested in assessing GTA TPD outcomes across programs and institutions collect data concerning other variables in the framework besides outcomes (e.g., GTA characteristic variables, implementation variables), as they might be important covariates. To the best of the authors’ knowledge, however, there are no known and broadly applicable instruments designed to elicit evidence of these other key categories of framework variables. The development of such instruments indeed constitutes a potential target of future scholarship. In particular, instruments could be designed to gather evidence concerning both GTA TPD contextual variables (i.e., institutional variables, GTA training design variables, and GTA characteristics) and implementation variables. These instruments could be administered to either TPD program staff or participating GTAs for data-collection purposes in the context of large-scale research.

---

2We refer the reader to Reeves and Marbach-Ad (2016) for information about how to select high-quality instruments.
Table 1. Possible instrumentation for collection of evidence concerning GTA TPD outcomes

<table>
<thead>
<tr>
<th>GTA TPD variable category</th>
<th>Specific GTA TPD program outcome variable</th>
<th>Example research/evaluation question</th>
<th>Possible categories of instrumentation</th>
<th>Example of existing instruments</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTA cognition</td>
<td>Knowledge/skills</td>
<td>Did participants acquire the intended knowledge and skills (in terms of pedagogy, assessment, and curriculum)?</td>
<td>Content tests; surveys</td>
<td>Pedagogy of Science Teaching Tests (Coburn et al., 2014)(^a)</td>
</tr>
<tr>
<td></td>
<td>Attitudes toward teaching</td>
<td>Was the GTA TPD associated with changes in participants’ valuing student-centered approaches?</td>
<td>Surveys; interviews</td>
<td>Survey of Teaching Beliefs and Practices (STEP; Marbach-Ad et al., 2014)</td>
</tr>
<tr>
<td></td>
<td>Beliefs about teaching</td>
<td>Was the GTAs’ teaching self-efficacy increased following the TPD?</td>
<td>Surveys</td>
<td>Science Teaching Efficacy Belief Instrument (Smolleck et al., 2006)</td>
</tr>
<tr>
<td>GTA teaching practices</td>
<td>Planning</td>
<td>Do GTAs who participated in the TPD use backward design to plan their classes?</td>
<td>Artifacts (e.g., lesson plans, assessments); surveys; interviews; focus groups</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Instruction</td>
<td>Do GTAs who participated in the TPD spend more time interacting with students?</td>
<td>Surveys; student evaluations of instruction</td>
<td>End-of-semester student evaluations used in Marbach-Ad et al. (2012)</td>
</tr>
<tr>
<td></td>
<td>Assessment</td>
<td>Following professional development, are GTA assessments more closely aligned to course learning outcomes?</td>
<td>Artifacts (e.g., assessments)</td>
<td>Rubric for examining objective-assessment alignment (Wyse et al., 2014)</td>
</tr>
<tr>
<td>Undergraduate student outcomes</td>
<td>Knowledge/skills</td>
<td>Do students taught by TPD-trained GTAs demonstrate improved knowledge and skills?</td>
<td>Content tests/concept inventories; surveys; interviews; artifacts (e.g., student work)</td>
<td>Test of Scientific Literacy Skills (Gormally et al., 2012); Genetics Concept Assessment (Smith et al., 2008)</td>
</tr>
<tr>
<td></td>
<td>Retention/attainment</td>
<td>Are students taught by GTAs who participated in TPD more likely to be retained in the biology major and graduate?</td>
<td>Official institutional and academic transcript data</td>
<td>Time to degree; first- to second-year retention; graduation</td>
</tr>
<tr>
<td></td>
<td>Interest</td>
<td>Do biology students taught by TPD-trained GTAs demonstrate greater interest in learning biology?</td>
<td>Surveys; interviews; focus groups</td>
<td>Colorado Learning Attitudes about Science Survey (Semsar et al., 2011)</td>
</tr>
</tbody>
</table>

\(^a\)For each of three general GTA TPD outcomes (i.e., GTA cognition, GTA teaching practice, and undergraduate student outcomes) and nine corresponding specific outcomes (e.g., GTA knowledge/skills, GTA planning, and undergraduate student retention), the table outlines example research/evaluation questions that might be asked by TPD program staff or researchers and possible categories of instrumentation and example specific instruments that might be used.

\(^b\)This instrument is intended for formative use in grades K–8 science teaching and is included for illustrative purposes only.

\(^c\)To the best of the authors’ knowledge, there are no currently published instruments to systematically elicit evidence of backward design planning, which is a fruitful area for future research.

CONCLUSION

The proposed conceptual framework explicated in this article was created with two purposes in mind: 1) to offer a guide for the evaluation of GTA TPD programs at individual institutions and 2) to offer a framework for how institutions can begin to coordinate evaluation and research efforts in order to build evidence-based biology GTA TPD practices.
Although we make no claims that the framework is comprehensive and complete, we believe that it can serve as a starting point for dialogue among practitioners and researchers about how to conduct large-scale, systemic research. The results generated from these coordinated efforts will, in turn, provide biology GTA TPD practitioners with empirical data that can be used to improve GTA teaching practices and undergraduate outcomes at their institutions.

For those who lead GTA TPD programs, we hope the conceptual framework provides insights to improve local programmatic evaluation practices. Program practitioners may realize, for example, that they have only been evaluating GTA satisfaction with their programs. In this case, they may use the information in this framework to begin to assess bonafide outcomes such as GTA cognition (e.g., knowledge of inquiry-based teaching methods). The conceptual framework could potentially be used as justification to department chairs or other administrators to provide additional resources to conduct these types of studies, particularly if the connection to undergraduate student outcomes is made clear.

The framework also provides practitioners with flexibility, a key factor given the multiple contexts in which biology GTA TPD is enacted. Practitioners may realize that they are only interested in probing the impact of GTA TPD enactment on only one particular outcome variable. Identifying the questions practitioners may wish to pursue and the resources they have available to pursue those questions will help them to build an evaluation plan that fits their particular needs. The example evaluation/research questions in Table 1 should guide those practitioners to identify specific questions and begin to think about the methods (instrumentation) they could use to assess them.

Finally, the conceptual framework proposes contextual variables that should be documented during dissemination of evaluation/research results for the purposes of more systematically comparing programmatic results across institutions. Ideally, researchers and practitioners at different institutions would coordinate their programmatic efforts as part of a designed research study, but we recognize that this may not be possible in practice because of the contextual variability in which programs at different institutions are enacted. Instead, collecting similar contextual variables and using some of the same instruments to measure program outcomes will allow institutions to compare their results and begin to hypothesize practices that may be beneficial at either particular types of institutions or at institutions more broadly. Comparisons such as these will greatly improve the ability of the field to move forward with identifying practices that maximize the impacts of TPD on GTAs and undergraduates (Schussler et al., 2015).

Given the profound impact that biology GTAs have on teaching at undergraduate institutions, enhancing GTA TPD as a means to improve GTA teaching practices and undergraduate learning outcomes should be a priority for institutions of higher education. Particularly for gateway science courses, improved GTA teaching practices may be a key lever to improve degree attainment in the sciences (e.g., O’Neal et al., 2007). As these GTAs move through their graduate programs, many will go on to become members of the professoriate; thus, providing effective biology GTA TPD programs may be one critical link to fully envisioning the promise of evidence-based teaching practices in biology courses.

REFERENCES


Austin A (2002). Preparing the next generation of faculty: graduate school as socialization to the academic career. J High Educ 73, 94–122.


Research Methods

A Primer for Developing Measures of Science Content Knowledge for Small-Scale Research and Instructional Use

Kristin M. Bass,*† Dina Drits-Esser,† and Louisa A. Stark‡

*† Rockman et al, San Francisco, CA 94105; ‡ Genetic Science Learning Center, University of Utah, Salt Lake City, UT 84102

Submitted July 5, 2015; Revised February 3, 2016; Accepted February 4, 2016

Monitoring Editor: Ross Nehm

The credibility of conclusions made about the effectiveness of educational interventions depends greatly on the quality of the assessments used to measure learning gains. This essay, intended for faculty involved in small-scale projects, courses, or educational research, provides a step-by-step guide to the process of developing, scoring, and validating high-quality content knowledge assessments. We illustrate our discussion with examples from our assessments of high school students’ understanding of concepts in cell biology and epigenetics. Throughout, we emphasize the iterative nature of the development process, the importance of creating instruments aligned to the learning goals of an intervention or curricula, and the importance of collaborating with other content and measurement specialists along the way.

INTRODUCTION

The quality of a scientific research project often depends on the availability of appropriately sensitive instruments. Advances in measurement, from nanoparticle biosensors (Howes et al., 2014) to disease diagnostics (Vogel, 2014) to digital PCR (Perkel, 2014), are critical for discoveries in basic and applied biological research. High-quality instruments are equally important in biology education. A well-conceived instrument that measures content knowledge is essential for making claims about the effectiveness of a new educational intervention and can provide data for curriculum or program improvement. It can be quite challenging, however, to identify instruments that are suitable for an intervention, to adapt existing items, or to create new ones altogether. Assessing the validity and reliability of these tools adds a level of complexity to the mix, especially within the budget constraints of small-scale projects.

Further, while science, technology, engineering, and mathematics faculty members are most familiar with assessment-related terms such as “student learning outcomes” and “summative assessment,” they are much less knowledgeable about issues related to the validity of interpretations of tests, surveys, or assessment items (Hanauer and Bauerle, 2015).1 They consequently write and use assessment items in courses they teach without undertaking a development process to determine how well their items measure the intended content. This approach may or may not be adequate for assigning student grades and is insufficient for studies that seek to determine the efficacy of new interventions. In fact, a recent analysis of evaluations of educational innovations in genetics and bioinformatics found that less than 10% of published articles contained any information about the reliability or validity of study instruments (Campbell and Nehm, 2013). Challenges may also arise when using published instruments or questions from item banks to evaluate learning. In these cases, the items may not be tightly aligned with the learning goals and objectives of courses, curricula, or other educational interventions.

1We define and discuss many common assessment terms, including reliability, validity, and analytical versus holistic scoring. Readers less familiar with core terminology may also wish to consult Rudner (1994) and the glossary in the Standards for Educational and Psychological Testing (AERA et al., 2014).
Further, if many of the items on an instrument are too easy or too difficult for the majority of study participants, little can be learned about the intervention’s efficacy.

This essay is an introduction to developing content knowledge assessments that will be used on a relatively small scale, as opposed to larger-scale assessment programs intended to reach a large audience (e.g., the National Assessment of Educational Progress [NAEP] or the Assessment of Higher Education Learning Outcomes) and that therefore require more advanced psychometric analyses. We, the authors, are two educational psychologists (K.B., D.D.-E.) and an evolutionary biologist (L.S.), all with extensive experience in program evaluation. The first two authors of the paper have worked closely together as external and internal evaluators for the University of Utah's Genetic Science Learning Center (GSLC) to carry out research studies on GSLC curricula and rigorous project evaluations on GSLC programs. We have geared this paper toward faculty who are involved in creating and evaluating curricula or other educational interventions (either for their own postsecondary courses or for a K–12 audience) and individuals who work with external program evaluators and wish to increase their understanding of instrumentation.

We first provide a brief overview of the process of instrument construction. We follow with a more detailed discussion of each step in the process, illustrated by examples from our assessments of high school students’ knowledge of cell biology and epigenetics (Drits-Esser et al., 2014). In particular, we emphasize how to create instruments aligned to the learning goals of an intervention or curricula and how to determine whether the items are valid for assessing the content they are intended to measure and in the contexts in which they will be used.

**A GENERAL INSTRUMENT-DEVELOPMENT PROCESS**

Measurement, whether in biology or education, involves making quantifiable inferences from observable evidence. Gel electrophoresis, for example, determines the relative length of DNA strands. It is not possible to see the strands directly; rather, biologists infer the length of strands based on the distance they migrate in the gel, as visualized by staining. Likewise, educators cannot see knowledge growth at any kind of direct, neurological level. Instead, they have to draw conclusions about what individuals know and can do through their performance on observable tasks. Crafting those tasks or items is both an art and a science, requiring detailed knowledge of the content being assessed, the population being measured, the contexts in which the instrument will be used, and the range of items or tasks that might appropriately elicit the content being tested.

A 2001 National Research Council (NRC) report entitled *Knowing What Students Know* and several reports and journals that followed (NRC, 2006, 2014; Songer and Ruiz-Primo, 2012) represent assessment design as a triangle with three vertices: cognition, observation, and interpretation (Figure 1). The cognition vertex refers to research on how students learn a topic and enables researchers or instructors to identify the precise “targets of inference” (NRC, 2001, p. 62) they wish to measure. The observation component of the triangle describes the tasks designed to draw out students’ knowledge and skills and differentiate students based on their levels of understanding and ability. The interpretation vertex refers to the reasoning and analytical tools used to make inferences about latent cognition from the gathered observations. This process utilizes statistical models with large-scale assessments but often has a qualitative component in smaller-scale settings (NRC, 2001, 2014).

The strength of the assessment triangle is that it concretely yet simply illustrates the notion of assessment as a way of reasoning from evidence (NRC, 2001, 2014). The alignment of research on student learning with observations of performance is made possible by interpretive tools and frameworks. In this model, assessment is not just referring to test items or formats, but denotes a process of making quality inferences about student ability using data from a carefully constructed sample of items or tasks (Campbell and Ném, 2013). It requires researchers or instructors to think explicitly about how they will use the data from the items they have written or selected to draw logical conclusions about what students have learned from a course or intervention.

The coordination of the triangle’s three elements is part of what makes developing assessments so difficult, especially in topic areas for which models of learning have not yet been fully specified. A special issue of the *Journal of Research in Science Teaching* contains several case studies of how researchers have attended to the cognition, observation, and inference vertices in their construction of instruments (Songer and Ruiz-Primo, 2012). We believe that it is also possible for novice developers to apply the assessment triangle to their work and describe a process for doing this.

Designing an instrument to measure content knowledge involves four basic steps: 1) identifying the concepts to be measured (also known as construct identification), 2) selecting or writing assessment items, 3) creating a scoring system, and 4) reviewing and validating items. The first step draws developers’ attention to the cognition component of the assessment triangle, while the second step focuses on methods of observing student knowledge and skills. The third and fourth steps address the interpretation of student performance, or the match between the observed evidence and the instrument’s intent. This multistep, construct-centered process is not entirely linear. Instrument developers frequently iterate between steps, for example, using the results of their validation studies to redesign their items or scoring, or using research on the constructs being assessed to inform the interpretation of student work (NRC, 2001). Throughout these steps, developers also need to attend to the connections between cognitive constructs, student observations, and
the interpretive methods that will enable them to make appropriate claims with their instruments. Figure 2 illustrates these four steps and lists some questions to consider during the development process.

As we expand on each of the four steps, we provide examples from our process of developing pre and post content knowledge assessments to measure high school biology students’ learning gains from two curriculum supplements developed by the GSLC. We used these instruments in small-scale randomized controlled trials that compared the GSLC-developed supplements with other materials that addressed the same learning goals (Drits-Esser et al., 2014).

The GSLC develops curriculum supplements that are freely available on its Learn.Genetics and Teach.Genetics websites. The materials include 1) interactive, multimedia learning experiences; 2) short movies; 3) three-dimensional animations; 4) “Learn More” web pages; 5) hands-on models; 6) paper-based learning activities; and 7) graphic organizers for students to use with the online materials. Each supplement addresses several broad learning goals with each learning experience typically focusing on one learning objective. Our primer discusses content measures for the *Amazing Cells* (GSLC, 2008a,b) and *Epigenetics* (GSLC, 2009a,b) curriculum supplements.

**Step 1. Construct Identification**

It is common to construct content knowledge assessments by reviewing the curriculum or program and diving straight into writing items that match the content of the individual learning activities. That approach misses a critical step. Reviewing a curriculum’s overall goals and objectives as well as applicable research on student learning first allows you to design a more cohesive set of questions that address the most important, enduring ideas.

A learning goal is an outcome, broadly stated, that students are expected to accomplish by the end of a unit, module, or intervention. An objective describes the specific knowledge and skills needed to achieve that goal. Objectives are commonly, though not always, phrased as action statements (e.g., explain, interpret, apply; Wiggins and McTighe, 2005). While there can be many objectives for a given goal, some may be more important to know than others. Useful resources for narrowing down which concepts to measure include curriculum developers, content experts, and state and national K–12 science frameworks and standards (e.g., College Board Standards for College Success [College Board, 2009], the Framework for K–12 Science Education [NRC, 2012], the Next Generation Science Standards [NGSS Lead States, 2013], the NAEP Science Framework [National...
Assessment Governing Board, 2015)), and undergraduate biology concept inventories (Garvin-Doxas et al., 2007).

In addition, before beginning a new unit or content area, it can be useful to assess the concepts students should already know and/or the misconceptions that may interfere with their learning (Sadler et al., 2012; NRC, 2014). Along with items about the content students are about to learn, this information provides a baseline from which to measure progress. If students do not achieve the intended learning objectives, these data can provide insights on why this may be the case.

As we planned our Amazing Cells and Epigenetics assessments, we worked with the GSLC curriculum developers to prioritize the learning objectives for the supplements. For example, the overarching goals of the Epigenetics supplement were for students to understand 1) what epigenetics is and 2) the relationship between epigenetics and the environment. There were six to eight more specific, measurable learning objectives within each goal. We asked the lead curriculum developer to differentiate the main ideas of the supplement from the ones that were simply “nice to know.” Within the two goals, she identified four key objectives for us to measure:

1. The epigenome influences gene expression.
2. Signals from the environment influence gene expression by acting on the epigenome. The epigenome helps cells “remember” the sum total of signals they have received that affect gene expression.
3. Epigenetics can lead to differences in genetically identical twins.
4. Factors from the environment such as diet, physical activity, and stress influence the epigenome.

It is often helpful to begin at the end and map backward from the ultimate goal of what is to be learned to students’ prior knowledge. As we articulated the objectives we would measure, we therefore considered the knowledge students had or would need to understand the epigenetics materials. High school students commonly assume that “a gene is a trait or that DNA produces proteins” (Elrod, 2007, p. 2). Students who approached the supplement materials with such misconceptions might have difficulty grasping the subsequent content. On the other hand, the materials might clear up some of those misconceptions, so we wanted to make sure to capture any potential learning gains. We consequently decided to assess students’ understanding of the relationship between DNA, genes, and proteins in addition to the supplement’s main objectives.

After narrowing down the objectives you want to measure (or even during this process), it is important to identify the kinds of observable evidence you will collect to demonstrate that the objectives have been met. This involves creating or selecting test items.

Step 2. Item Parameters, Selection, and Creation

Item Parameters. There are many ways to assess a given objective. We will discuss two of the most common item formats in content assessments: multiple-choice (MC) and open-ended items (also known as constructed responses, such as short answer, essays, and fill-in-the-blank). These formats have several inherent advantages and drawbacks. For example, MC items are straightforward to score and employ a response format that is quite familiar to most students (Simkin and Kuechler, 2005). However, while MC items can be constructed to access a range of levels of students’ cognition, they also have certain limitations, including the difficulty and time investment of developing quality items, potential false indication of students’ knowledge and understanding, and potential demographic biases associated with performance (e.g., gender, ethnicity, socioeconomic status). Further, these questions may limit the potential to assess students’ ability to think creatively and to organize and synthesize information (Simkin and Kuechler, 2005) and may perpetuate the idea that scientific investigation has only one correct answer. Open-ended items provide for more nuance and variability in student responses, which can access divergence in students’ thought processes, and are more likely to capture conceptual learning (Martínez, 1999). However, grading may be time intensive, subjective, and highly sensitive to variation across raters (Simkin and Kuechler, 2005; Stanger-Hall, 2012).

Other options for measuring science knowledge include performance-based assessments (i.e., hands-on activities requiring students to conduct all or part of a science investigation to demonstrate their knowledge and skills [NRC, 2014]), problem scenarios posed within educational computer games (Hickey et al., 2009), or analysis of students’ science laboratory notebooks (Baxter et al., 2001; Wallert and Provost, 2014). In some cases, these other types of assessments are more difficult to design and to score; however, they may access students’ cognition in ways that MC or short-essay items cannot. We nevertheless focus on these two latter question formats in this article, because they are most familiar to faculty and students and therefore appropriate for this primer.

Before we began drafting instruments for the Amazing Cells and Epigenetics supplements, we first had to set some parameters for the number of items we needed. We had to be considerate of the amount of time teachers had to administer the instrument, the length of the intervention, and the age of the participating students. Further, we decided that we generally wanted at least three items per objective; three is the minimum number of replications recommended in scientific measurement, and we felt the same standards were suitable here.

We ultimately decided that the instruments should take students approximately 15–20 min to complete, given that we were evaluating high school students’ performance on a two-day intervention. Assuming approximately 1 min per MC item and at least 5–7 min for essays, we planned for roughly 12–18 items divided equally among assessment objectives. We also knew that, to arrive at our target number of items, we would have to come up with at least twice as many at the beginning of our process; experience had shown that we would probably eliminate half of our items during internal and external review processes (for instance, because an item did not address the most important content in the curriculum supplement).

It is also important to consider the relative difficulty of the items you wish to develop. Asking questions that require a range of thinking skills will help you better evaluate where students fall on a continuum of learning for each objective. For example, the Trends in International Math and Science Study (TIMMS; Mullis et al., 2009), a survey of students in the elementary and secondary grades, classifies items into
three cognitive domains: knowing (recall and vocabulary), applying (making connections between concepts), and reasoning (using content and process knowledge to solve problems and construct explanations). You may wish to include similar kinds of lower- and higher-order thinking items in your content knowledge instruments. Keep in mind that item format is not necessarily related to cognitive domain, nor is item difficulty necessarily related to either item format or cognitive domain. MC questions can assess reasoning as well as recall, while essays and performance-based assessments may inadvertently elicit knowledge of facts instead of their application (Baxter and Glaser, 1998; NRC, 2014).

**Item Selection and Adaptation.** In selecting our items, we first reviewed existing instrument databases, including the ones listed in Table 1. We also looked at the Genetics Literacy Assessment Instrument (GLAI; Bowling et al., 2008), a concept inventory intended for undergraduate non-science majors, which we thought might be suitable for the Epigenetics supplement. Several of the items addressed the objectives we intended to measure, but the reading level was too difficult for high school students. We knew that, by simplifying some of the language, we would make the items more valid for measuring our particular population. We also recognized that, if we altered the items, we could not claim that our instrument had the same reliability and validity characteristics of the GLAI as published, nor would we be able to compare the performance of our students with that of students in other studies conducted with that instrument. Because the purpose of our instrument was to evaluate the effectiveness of the GSLC’s curricular materials, we decided that suitability for the student population being assessed outweighed any "interest in generalizing to other groups."

We adjusted the items mainly by shortening sentences or phrases. Table 2 displays one of the original GLAI items and our adaptation. We simplified the phrase “expression of his or her genes” to “gene expression,” but expanded the phrase “lasting until adulthood” to “stopping when a person reaches adulthood.” We added italics to focus attention on the key differences between answer choices. Additionally, we looked for any terms that high school students might not understand and that would not be addressed in the Epigenetics materials. For instance, in this item, we removed the term “menopause,” since we thought it might not be familiar to some high school students, and it was not essential to understanding the concept we intended to test. We also divided the question into two parts, the first of which required a yes or no answer. Our desire to retain the integrity of the original GLAI item inadvertently made the single no option much more conspicuous. It is generally advisable to have even numbers of options with a particular stem, so test-takers cannot easily eliminate a choice. We will discuss this issue in further detail in the next section. Nevertheless, we have chosen to share this item to highlight some of the trade-offs encountered with revising items and the need for as many iterations of development, discussion, and revision as time will allow.

**Item Creation.** We often could not find enough existing items for an objective and had to create our own. In these cases, we looked to the curriculum supplements for inspiration and drafted items in accordance with recommended item-writing guidelines, such as keeping the text simple to minimize the time needed to read items, using positive—not negative—phrasing, and avoiding options such as “I don’t know” (Haladyna et al., 2002; American Association for the

### Table 1. Examples of publicly available instrument databases with life science items

<table>
<thead>
<tr>
<th>K–12</th>
<th>Higher education</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trends in International Mathematics and Science Study (TIMSS): <a href="http://www.timss.org">www.timss.org</a></td>
<td></td>
</tr>
</tbody>
</table>

**Table 2. Adaptation of a Genetics Literacy Assessment Instrument item**

<table>
<thead>
<tr>
<th>Original</th>
<th>Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>At what times during an individual’s life does the environment influence the expression of his or her genes?</td>
<td>Can the environment influence gene expression? If so, during which times in an individual’s life?</td>
</tr>
<tr>
<td>A. Beginning at conception and lasting throughout life.</td>
<td>A. Yes, beginning at conception and lasting throughout life.</td>
</tr>
<tr>
<td>B. Beginning at birth and lasting throughout life.</td>
<td>B. Yes, beginning at birth and lasting throughout life.</td>
</tr>
<tr>
<td>C. Beginning at birth and lasting until adulthood.</td>
<td>C. Yes, beginning at birth and stopping when a person reaches adulthood.</td>
</tr>
<tr>
<td>D. Occurring only during key stages of life such as puberty and menopause.</td>
<td>D. Yes, but only during key stages of life such as puberty.</td>
</tr>
<tr>
<td>E. Environment has little or no effect on how genes are expressed.</td>
<td>E. No, the environment has little or no effect on gene expression.</td>
</tr>
</tbody>
</table>

Correct answer: “A.”
Table 3. Tips for writing MC questions

General
- Test for important or significant information (base each question on student learning objective of the lesson, not trivial information).
- Be sure the item would be comprehensible to your students.
  - Avoid unfamiliar vocabulary that is not defined and that is not related to the learning goal.
  - Avoid complex sentences.
  - Avoid words and phrases with confusing or ambiguous meanings.
- Items should have only one right answer.
- Use present tense and active voice.
- Minimize the time required to read each question.

Stem
- Include the central idea to avoid repetition in answer choices.
- Keep sentences brief and straightforward with a simple phrase structure and no additional clauses.
- Word positively—avoid negative phrasing.
  - Avoid phrasing “all of the following except” or “which of the following is false.”

Answer choices
- Link one or more of the distractors to misconceptions related to the key idea.
- Each answer choice should be a single word or phrase or a single sentence (keep options short).
- Keep all options homogeneous in content.
- Keep answer choice length similar.
  - Avoid “all of the above.”
  - Avoid “none of the above.”
  - Avoid “I don’t know.”
- Include from three to five options for each question.
- Keep options independent; options should not be overlapping.
- Phrase options positively, not negatively.
  - Avoid distractors that can clue test-wise examinees (e.g., absurd options, formal prompts, or overly specific or overly general clues).
  - Avoid giving clues through the use of faulty grammatical construction.

Illustrations
- Keep illustrations simple and to the point.
- Illustrations should facilitate the understanding of what is being asked.
- Include the same information in the text and the illustration.

Cytokine is a type of cell signal. How might a liver cell respond to cytokine?
A. Probably in different ways depending on how far the cytokine travels.
B. Probably in the same way that a blood cell will respond to cytokine.
C. Probably in a different way than a blood cell will respond to cytokine.
D. Probably in the same way the liver cell would respond to a nitric oxide cell signal.

Figure 3. Sample assessment item written for the Amazing Cells supplement (correct answer: “C”).
"same way" and "different way" options, it might have cued test-savvy students to eliminate the choice that did not resemble the others.

A complementary aspect of designing items is determining how to quantify students’ responses. This is discussed next.

**Step 3. Item-Scoring Systems**

Different from a simple answer key, a scoring system establishes clear, consistent criteria for the number of points to be assigned to responses of varying quality. This section describes some strategies for creating rubrics for open-ended written responses. Keep in mind, however, that it is possible to give full or partial credit to any type of item, including MC questions (Briggs et al., 2006; Hadenfeldt et al., 2013).

Rubric development should happen concurrently with item construction. When drafting an open-ended item, you should establish criteria for a complete and correct answer, and speculate on the kinds of responses that would demonstrate partial understanding of the targeted idea. The levels of a rubric, or the scores that are assigned, can be refined by reviewing and categorizing actual student work. This “top-down, bottom-up” process (Chi, 1997) ensures that the final rubric accounts for a priori expectations for what an item should measure while being sensitive to the realities of student performance.

Designers must make a number of decisions to produce a quality rubric. One choice is establishing the number of criteria upon which to rate responses. It may be possible to score a response using a holistic rubric that evaluates the overall quality of an answer. An analytical rubric, which rates performance on several components (e.g., the quality of claim, evidence and reasoning in a scientific argument; McNeill and Krajcik, 2011), requires more time to develop and use but permits greater precision in measuring knowledge and skills.

A related rubric design issue is identifying the most important content for a response that will receive high scores (as opposed to ideas that would be useful for a student to include, but not necessary; Arter and McTighe, 2001). It can therefore be beneficial to revisit the specific goals and objectives you wish to assess and ask how well each detail in a rubric addresses those objectives. Rubric creators must also justify the inclusion of quantitative and qualitative criteria for judging performance. The number of examples or pieces of evidence in a response may not be as important as the quality of those examples. It is important to differentiate what can be counted from what actually “counts.”

Finally, a rubric should be so clear and comprehensive that all raters can use it to agree on scores. The best way to evaluate this is to allocate enough time in the assessment-development process to train raters and establish interrater reliability or consistency. It is usually easier to get good interrater reliability using analytical rubrics than holistic ones, so that is another factor to take into consideration when deciding which type to use. When pilot testing a rubric for short open-ended responses, we recommend that raters score a random sample of 10–20 answers and then compare scores. You may choose to establish exact agreement or allow some degree of variation depending on the number of categories in the rubric, the time and resources available for scoring, and the magnitude of the consequences of the scoring decisions (e.g., high-stakes college admissions or teacher promotion; Stemler and Tsai, 2008). There are also different methods for calculating interrater reliability, ranging from percentage of agreement (the most intuitive to interpret) to Pearson's $r$ correlations, Cohen's kappa, and intraclass correlation coefficients (ICCs). Interested readers can consult Stemler and Tsai (2008) and Hallgren (2012) for formulas, along with SPSS and R syntax for these calculations. Rather than explaining the merits and limitations of each of these statistics, we wish to emphasize the general value of using multiple raters to inform instrument design.

Discussions about scoring disagreements can lead to changes in rater training or a clarification of rubric criteria, as we will illustrate in an example from the *Epigenetics* assessment.

As part of developing holistic rubrics for open-ended *Epigenetics* items, we conducted a pilot test of the entire assessment with 83 high school students who had just completed the GSLC *Epigenetics* supplement. This allowed us to determine whether the items were eliciting the answers we expected (a point on which we will elaborate in the next section), and helped us generate initial ideas for the scoring rubric.

One of the questions we piloted was, "What does the epigenome do?," which, while broad, related directly to the goals and objectives of the supplement. We expected that students would be able to explain that the epigenome controls gene expression by turning genes on and off. We assumed that we would have at least two categories of partially correct responses but were not sure what the criteria would be for those levels. As we analyzed the pilot data, we determined that students’ partially correct answers generally fell into one of four themes or categories (Table 4).

<table>
<thead>
<tr>
<th>Rubric category</th>
<th>Examples from students’ responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>The epigenome is involved in gene expression.</td>
<td>The genome is the full genetic information of a human. The epigenome is what tells those genes how to be expressed.</td>
</tr>
<tr>
<td>The epigenome reacts to the environment.</td>
<td>The epigenome is sitting above the genome and changes chemical signals according to environmental factors.</td>
</tr>
<tr>
<td>The epigenome influences traits.</td>
<td>The epigenome decides what traits you get.</td>
</tr>
<tr>
<td>The epigenome reacts to signals.</td>
<td>The epigenome arranges in response to signals.</td>
</tr>
</tbody>
</table>

3Reliability refers to the ability of an instrument to produce the same results in different situations. It can be assessed with different types of data, each related to a particular source of error or inconsistence in measurement. In this paper, we discuss strategies for establishing agreement between raters. Reliability can also refer to the consistency of items to measure the same underlying idea (internal consistency) or of an instrument to generate similar responses over time (test–retest reliability) (Cook and Beckman, 2006; Lovelace and Brickman, 2013). A consideration of these issues is beyond the scope of this article.
In internal memos, we asked ourselves, “Are all of the partially correct categories the same quality, or are some more correct than others?” We decided that two of the categories demonstrated a more accurate, albeit still incomplete, understanding of the supplement’s objectives than the other two categories. We therefore divided the partially correct answers into two levels. Level 1 consisted of answers about the epigenome influencing traits and reacting to signals, and Level 2 contained answers mentioning that the epigenome is involved in gene expression or reacts to the environment. For the curriculum field test, we developed a four-level rubric giving three points to responses that completely and correctly answered the question, one or two points for answers that met the partially correct criteria we had established, and zero points to answers with significant misconceptions.

We continued to adjust the partially correct rubric categories through conversations about our ratings. For instance, we initially disagreed on how to rate responses that indicated only that “the epigenome controls the genome.” One rater wanted to give these answers a zero, because they repeated parts of the question, while the other rater thought the answers deserved a one, because they were more accurate than the other answers she had scored as zeros. We ultimately accepted the latter rater’s argument and modified the rubric accordingly. Once we felt confident about our rubric, we assigned one person to score all of the responses to that question from our field test and another to score a random 25% sample. We then calculated ICCs to evaluate interrater reliability (Drits-Esser et al., 2014).

For the Epigenetics assessment, we codeveloped our rubrics jointly and then evaluated interrater reliability. Alternatively, a researcher may construct a rubric independently and train another individual to serve as a second rater. Rubric clarity is particularly important in these cases, since the rater may be less familiar with the assessment context and the data at hand. An effective rubric should contain definitions of each level that specify the depth of understanding and even the terminology a student should provide to receive credit for that level (Allen and Tanner, 2006). Rubrics should also contain two or more examples per level to reinforce the definitions (Arter and McTighe, 2001).

Even the most detailed rubrics may not be able to accommodate all possible responses. Borderline cases are inevitable. We recommend reviewing some ambiguous answers during rater training and justifying why each response should fall into one level or another. Such conversations may cause you to further clarify the difference between levels (Moskal and Leydens, 2000). If there is a sufficient number of borderline answers with common characteristics (e.g., 10% of your sample or more), you may wish to create a new rubric level entirely.

Finally, keep in mind that while we have been discussing interrater reliability to this point, intrarater reliability is equally important. It is good practice to rescore a small sample (10%) of your own ratings to ensure you have been consistent all the way through. It is common to experience “rater drift” and become more lenient or stringent over time; rescoring some answers or comparing the first answers you scored with the last helps you determine whether you need to adjust some of your ratings.

Once you have assessment items, it is time to validate them. This process utilizes several lines of evidence to build a credible argument about the appropriateness of your items for the context in which you will use them.

**Step 4. Item Review and Validation**

Validity is defined by the Standards for Educational and Psychological Testing as “the degree to which accumulated evidence and theory support a specific interpretation of test scores for a given use of a test” (American Educational Research Association [AERA] et al., 2014, p. 225). In other words, how well are you measuring what you intended to measure? To what extent can you justifiably use instrument scores to infer what students know at a particular time point or evaluate the efficacy of an intervention?

Validity is not a property of a test but of its interpretation. While many granting agencies require the use of “valid and reliable” instruments, that request is a bit misleading. An instrument may be valid for one context but not another, much like a drug may be indicated for one condition or population but lacks evidence from clinical trials to support other uses. As noted earlier, we commonly encounter this issue when we use or adapt items from concept inventories or other instruments that have been validated for college undergraduates but not younger students.

It would be convenient to think about validity as a single quantifiable indicator, much like information about reliability is reported as correlations or alpha coefficients. However, validity is not a numerical rating per se, but rather an argument that can be supported with different kinds of evidence (Kane, 2013). It is helpful to conceptualize validity as a way to empirically test the claims you are making about your instrument. Think back to your original plans and remind yourself of what you wanted to measure and why. Then collect the most feasible and convincing evidence you can (within your time and budget constraints) to support your interpretations of student knowledge, based on the data you have gathered from your items. In this manner, validation brings the instrument-development process full circle.

If validation is represented as an argumentation process, the question then becomes what sources of evidence might be appropriate to justify various claims. The Standards for Educational and Psychological Testing (AERA et al., 2014) identify five sources: 1) test content, 2) response processes, 3) internal structure, 4) relationships to other variables (a.k.a. external structure), and 5) consequences. Note that it is not necessary to collect data from all five categories. Rather, the amount or type of evidence you need depends on the stakes attached to the instrument and the breadth of its use, as well as the availability of relevant data. We will discuss two types of validity evidence: *test content*, which considers how well items represent a concept or domain; and *response processes*, which examine the reasoning students apply to test answers. We determined that these sources were the most important for our early-stage pilot work, since they could directly inform item revisions.

It is also possible to validate assessments using evidence based on *internal structure* (i.e., the associations between test items, as they relate to the measure’s intended constructs), *relationships with other variables* (i.e., the relationship between students’ scores on the assessment you have developed to performance on established measures of similar or different constructs), and *consequences* (i.e., the effect of test scores on
positive social outcomes such as improvements in teaching practice or negative outcomes such as cheating). Readers are encouraged to consult Campbell and Nehm (2013), Reeves and Marbach-Ad (2016), and the Testing Standards for a more extended discussion of validity and suggested methodologies for gathering each type of evidence.

Validation Evidence Based on Test Content. The wording, content coverage, and format of a test and the conditions for its administration and scoring are all considered elements of test content (AERA et al., 2014). These should be reviewed internally and externally throughout the assessment design process. We routinely sent drafts of our Amazing Cells and Epigenetics items to the science content specialists and educators at the GSLC who had developed the two curriculum supplements. For instance, we sent a memo listing several questions we wanted the Amazing Cells developers to consider in their review:

- Do the items accurately represent the content covered in the Amazing Cells learning objectives? Are there any items in which the content is incorrect or the answer choices are confusing?
- Do the items represent the cognitive skills required for the unit (e.g., memorization, application, synthesis)? Do any of the items seem to be too easy or too hard?
- How well do you think students will be able to answer these questions—will they understand this metaphor or that vocabulary word? Is this question too long, or that one too short?

The reviewers responded with comments about the accuracy of the content and clarifications on the supplement’s goals. One of the objectives for Amazing Cells was that “cells communicate by sending and receiving signals.” One of the educators noted that her “intent was to depict signals as only being chemical.” We consequently constructed more items that addressed the general function of cell signals (i.e., to send messages inside the body) rather than representing signals as one specific form or another.

While we determined that the supplement developers would be the most appropriate individuals to review our items, faculty and others with expertise in a particular field are an obvious resource for external review (Kalas et al., 2013; Perez et al., 2013; Deane et al., 2014). Perez et al. (2013), for example, asked reviewers, “Is the correct answer accurate given the scenario?” and “Do any of the other answers strike you as correct?” (p. 671). K–12 teachers or science specialists may also be recruited to evaluate the appropriateness of item content and wording for the students they teach (Fives et al., 2014). They may point out, for example, vocabulary that their students may not understand or scenarios with which students may not be familiar. It is also a good idea to get students’ perspectives on items directly, as we will explain in the next section.

Validation Evidence Based on Response Processes. Working directly with the intended test takers helps assessment designers understand whether items actually require the content knowledge and reasoning skills they are supposed to elicit. Students may be able to answer an MC item correctly or eliminate distractors because of their familiarity with general test-taking strategies but not the content being tested. On the other hand, students may know the content but answer a question incorrectly because they apply additional information to their response.

A good example of the latter point comes from an evaluation of an NAEP hands-on performance assessment (Bass et al., 2002). In that study, interviewers prompted fourth-grade students to talk about what they were thinking as they conducted an experiment on sinking and floating. Students observed how high a pencil floated in freshwater and a salt solution, then repeated the procedure to identify the composition of an unknown “mystery water.” At the end of the activity, students answered a question designed to test their ability to apply what they had observed: “When people are swimming, is it easier for them to stay afloat in the ocean or in a freshwater lake? Explain your answer” (O’Sullivan et al., 1997, p. 45). Much to the interviewers’ surprise, some students—who had stated moments before that a pencil floats higher in saltwater than fresh—said that it was easier to float in a lake. Their rationale? The ocean has waves and sharks, not to mention the fact that salt water stings when it gets in your nose or eyes. It can be awfully hard to float in such challenging conditions. Researchers interpreted these responses to mean that even though students had the knowledge they needed to answer the question, they were considering other factors in their response besides the concentration of salt in a body of water (Bass et al., 2002). While this example is humorous, it illustrates the serious point that test takers can select answer items for very different reasons than you might expect. It is always a good idea to pilot test your assessments to uncover any unintended interpretations.

In our projects with the GSLC, we used two different strategies for examining students’ response processes. For Amazing Cells, we used a technique called a cognitive interview with three 10th-grade students who had not studied cell biology and one 11th-grade student who had studied cell biology. In a cognitive interview, respondents are asked to share what they are thinking as they answer a test item. They may also be asked follow-up questions about their answers (Hamilton et al., 1997). We suggest intentionally selecting students from diverse backgrounds and differing achievement levels for cognitive interviews, since this will provide useful information in adjusting items and removing jargon. Further, all students should have at least some knowledge of the content being tested. This will allow you to understand the types of responses that the items will evoke for a typical test taker in the population of interest.

We identified 11 MC questions that we had drafted for which we wanted feedback. Many of these items contained some scientific terminology that was not covered in the supplement or used metaphors for cell structure and function that we wanted to vet. We used a standardized cognitive interview protocol (Horizon Research, 2009), which asked students why they chose a particular answer and what they thought of the other answer choices. We followed up with questions about whether any of the words or visuals were confusing or might confuse other students. The latter question is especially valuable, because it enables students to “save face” if they do not know something.

We eliminated or edited several questions based on our interviews. We found that a student answered one question
about cell communication disruption (Table 5) correctly because she had recognized a pattern: “most of the questions so far have been on cell communication,” leading her to the decision that diseases such as diabetes and multiple sclerosis must be caused by breakdowns in cell communication. We chose to retain this question for the final assessment, but placed it at the beginning of the test so that students could not use information from the other items to inform their response. We also changed one of the distractors from “cell differentiation” to “cell division,” because some students were not familiar with the former term and eliminated that choice simply because they did not know what it was.

The second strategy we used to examine students’ responses to the items was classroom-level pilot testing, in which students completed a full-length test and provided feedback about items they did not understand. Compared with cognitive interviews, which we conducted with students individually, classroom-level testing allowed us to obtain information from a larger number of students in a relatively short amount of time. In planning our pilot tests, we had to make several decisions about the type of information we wanted to extract from the data. We had to determine the number of students needed to obtain reliable data, and we had to determine the level of student exposure to the content in school before the pilot test. We also needed to decide on the most appropriate student demographics for our needs (e.g., grade, ethnicity, gender, percent free or reduced lunch).

For Amazing Cells, we conducted the pilot test with 79 students in three 10th-grade biology classrooms, using items we had refined with data from the cognitive interviews. We piloted with students who had not previously studied cell biology in school in order to identify any items that could be answered with little or no prior knowledge about the topic; these items were eliminated from our final assessments. For Epigenetics, we chose to test with 83 biology students from three classrooms who had just completed the GSLC Epigenetics supplement in order to identify and eliminate items that were too difficult (i.e., that the majority of students could not answer correctly even after exposure to the material). We determined that three classes’ worth of data should produce adequate variation in response patterns while keeping the qualitative data to a manageable level. Additionally, our priority was testing with the appropriate grade level rather than testing with students of certain ethnicities or other demographic indicators. If we had been interested in identifying items that might be biased toward different groups, we might have increased our sample size and performed specialized analyses such as differential item functioning (Ostler and Everson, 2009).

Knowing that we would refine our items based on data from the pilot—or eliminate questions entirely—we administered tests with 23 items for Amazing Cells and 12 items for Epigenetics, more than we expected to appear on the final tests (the final Amazing Cells tests had 16 MC and open-ended items, and the final Epigenetics tests had 8 MC and open-ended items). We applied basic item analyses to students’ responses. To evaluate item difficulty, for instance, we ran frequencies for the percentage of correct answers and for each incorrect distractor students selected. It is generally recommended to use items that 30–80% of respondents answer correctly, especially if you hope to measure change from pretest to posttest (Kehoe, 1995). We could also have examined item discrimination statistics, which compare a student’s performance on a single item with his or her total test score. In other words, does a student who gets a particular item correct also score high on the rest of the test, or is it possible to get an item correct but score poorly on the other questions? A correlation of 0.15 or less indicates that an item does not effectively delineate between high- and low-performing students and should be eliminated (Kehoe, 1995).

Both item difficulty and item discrimination statistics can be calculated with most standard statistics packages or spreadsheet tools and are therefore recommended for simple item analyses. Researchers or instructors constructing high-stakes measures administered to large numbers of students may wish to consult with a psychometrician to perform more complex analyses utilizing Rasch modeling and generalizability theory. Rasch modeling enables test developers to estimate the difficulty of different items on the same continuous scale and construct multiple assessment forms (Wilson, 2005; Bond and Fox, 2007), while generalizability theory can be used to estimate the number of items, raters, and testing occasions needed to minimize measurement error and obtain an optimally reliable estimate of performance (Shavelson and Webb, 1991).

Table 5 provides examples of two Epigenetics items we refined based on feedback from the pilot test. In example 1, we replaced one item about the relationship between DNA, traits, and proteins with another item about the function of DNA. Only 24.7% of the pilot students answered this item correctly, slightly under our threshold of 30%. We had expected a higher percentage of correct responses, given that students had just completed the Epigenetics supplement. On further reflection, however, we decided that the concept of traits was not heavily emphasized in the supplement. This fact, compounded with the complexity of the answer choices (each of which compared three different terms), led us to generate another, simpler item to assess students’ understanding of the central dogma.

Example 2 illustrates our revision of the open-ended question we discussed in the previous scoring section. We not only developed a rubric that was sensitive to students’ responses, but adjusted the item to better elicit the information we wanted. We realized from the responses that we were more interested in knowing about what the epigenome does, which is a higher-level concept, rather than simply what it is. We revised the question accordingly.

### Table 5. Revision of a cell communication item based on cognitive interview results (change in italics)

<table>
<thead>
<tr>
<th>Original</th>
<th>Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diseases such as diabetes and multiple sclerosis occur when there is a breakdown in:</td>
<td>Diseases such as diabetes and multiple sclerosis occur when there is a breakdown in:</td>
</tr>
<tr>
<td>A. cell communication</td>
<td>A. cell communication</td>
</tr>
<tr>
<td>B. cell motion</td>
<td>B. cell motion</td>
</tr>
<tr>
<td>C. cell division</td>
<td>C. cell division</td>
</tr>
<tr>
<td>D. cell differentiation</td>
<td>D. cell growth</td>
</tr>
</tbody>
</table>

Correct answer: “A.”

For Amazing Cells, we conducted the pilot test with 79 students who had not previously studied cell communication, leading her to the decision that diseases such as diabetes and multiple sclerosis might be caused by breakdowns in cell communication. We chose to retain this question for the final assessment, but placed it at the beginning of the test so that students could not use information from the other items to inform their response. We also changed one of the distractors from “cell differentiation” to “cell division,” because some students were not familiar with the former term and eliminated that choice simply because they did not know what it was.

The second strategy we used to examine students’ responses to the items was classroom-level pilot testing, in which students completed a full-length test and provided feedback about items they did not understand. Compared with cognitive interviews, which we conducted with students individually, classroom-level testing allowed us to obtain information from a larger number of students in a relatively short amount of time. In planning our pilot tests, we had to make several decisions about the type of information we wanted to extract from the data. We had to determine the number of students needed to obtain reliable data, and we had to determine the level of student exposure to the content in school before the pilot test. We also needed to decide on the most appropriate student demographics for our needs (e.g., grade, ethnicity, gender, percent free or reduced lunch).

For Amazing Cells, we conducted the pilot test with 79 students in three 10th-grade biology classrooms, using items we had refined with data from the cognitive interviews. We piloted with students who had not previously studied cell biology in school in order to identify any items that could be answered with little or no prior knowledge about the topic; these items were eliminated from our final assessments. For Epigenetics, we chose to test with 83 biology students from three classrooms who had just completed the GSLC Epigenetics supplement in order to identify and eliminate items that were too difficult (i.e., that the majority of students could not answer correctly even after exposure to the material). We determined that three classes’ worth of data should produce adequate variation in response patterns while keeping the qualitative data to a manageable level. Additionally, our priority was testing with the appropriate grade level rather than testing with students of certain ethnicities or other demographic indicators. If we had been interested in identifying items that might be biased toward different groups, we

 Knowing that we would refine our items based on data from the pilot—or eliminate questions entirely—we administered tests with 23 items for Amazing Cells and 12 items for Epigenetics, more than we expected to appear on the final tests (the final Amazing Cells tests had 16 MC and open-ended items, and the final Epigenetics tests had 8 MC and open-ended items). We applied basic item analyses to students’ responses. To evaluate item difficulty, for instance, we ran frequencies for the percentage of correct answers and for each incorrect distractor students selected. It is generally recommended to use items that 30–80% of respondents answer correctly, especially if you hope to measure change from pretest to posttest (Kehoe, 1995). We could also have examined item discrimination statistics, which compare a student’s performance on a single item with his or her total test score. In other words, does a student who gets a particular item correct also score high on the rest of the test, or is it possible to get an item correct but score poorly on the other questions? A correlation of 0.15 or less indicates that an item does not effectively delineate between high- and low-performing students and should be eliminated (Kehoe, 1995).

Both item difficulty and item discrimination statistics can be calculated with most standard statistics packages or spreadsheet tools and are therefore recommended for simple item analyses. Researchers or instructors constructing high-stakes measures administered to large numbers of students may wish to consult with a psychometrician to perform more complex analyses utilizing Rasch modeling and generalizability theory. Rasch modeling enables test developers to estimate the difficulty of different items on the same continuous scale and construct multiple assessment forms (Wilson, 2005; Bond and Fox, 2007), while generalizability theory can be used to estimate the number of items, raters, and testing occasions needed to minimize measurement error and obtain an optimally reliable estimate of performance (Shavelson and Webb, 1991).

Table 5 provides examples of two Epigenetics items we refined based on feedback from the pilot test. In example 1, we replaced one item about the relationship between DNA, traits, and proteins with another item about the function of DNA. Only 24.7% of the pilot students answered this item correctly, slightly under our threshold of 30%. We had expected a higher percentage of correct responses, given that students had just completed the Epigenetics supplement. On further reflection, however, we decided that the concept of traits was not heavily emphasized in the supplement. This fact, compounded with the complexity of the answer choices (each of which compared three different terms), led us to generate another, simpler item to assess students’ understanding of the central dogma.

Example 2 illustrates our revision of the open-ended question we discussed in the previous scoring section. We not only developed a rubric that was sensitive to students’ responses, but adjusted the item to better elicit the information we wanted. We realized from the responses that we were more interested in knowing about what the epigenome does, which is a higher-level concept, rather than simply what it is. We revised the question accordingly.

### Table 5. Revision of a cell communication item based on cognitive interview results (change in italics)

<table>
<thead>
<tr>
<th>Original</th>
<th>Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diseases such as diabetes and</td>
<td>Diseases such as diabetes and</td>
</tr>
<tr>
<td>multiple sclerosis occur when</td>
<td>multiple sclerosis occur when</td>
</tr>
<tr>
<td>there is a breakdown in:</td>
<td>there is a breakdown in:</td>
</tr>
<tr>
<td>A. cell communication</td>
<td>A. cell communication</td>
</tr>
<tr>
<td>B. cell motion</td>
<td>B. cell motion</td>
</tr>
<tr>
<td>C. cell division</td>
<td>C. cell division</td>
</tr>
<tr>
<td>D. cell differentiation</td>
<td>D. cell growth</td>
</tr>
</tbody>
</table>

Correct answer: “A.”
Table 6. Revision of two Epigenetics items based on classroom pilot test results

<table>
<thead>
<tr>
<th>Pilot test</th>
<th>Final test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. What is the relationship between genes and traits? a</td>
<td>1. Which of the following statements about DNA or genes is the most</td>
</tr>
<tr>
<td>A. Genes code for DNA. DNA is responsible for individual traits.</td>
<td>accurate?</td>
</tr>
<tr>
<td>B. Genes code for proteins. Proteins are responsible for individual traits.</td>
<td>A. DNA provides the instructions for making proteins.</td>
</tr>
<tr>
<td>C. Genes code for chromosomes. Chromosomes are responsible for individual traits.</td>
<td>B. Genes provide the instructions for making DNA.</td>
</tr>
<tr>
<td>D. Genes code for carbohydrates. Carbohydrates are responsible for individual traits.</td>
<td>C. DNA provides the instructions for making carbohydrates.</td>
</tr>
<tr>
<td>E. Genes are not related to traits. The environment is primarily responsible for individual traits.</td>
<td>D. Carbohydrates provide the instructions for making DNA.</td>
</tr>
<tr>
<td>2. What does the epigenome do?</td>
<td>2. Explain (a) what the epigenome does to the genome, and (b) how it does it.</td>
</tr>
</tbody>
</table>

a Correct answers for item 1: pilot: “B”; final: “A.”
b Adapted from the Genetics Literacy Assessment instrument (Bowling et al., 2008).

After we had used the feedback from the cognitive interviews and classroom pilot tests to revise individual items, we still had to assemble those items into the final pre and post assessments to be used in the curriculum field test. In this final step, we again used strategic placement of items, careful alignment to the supplement objectives, and adherence to time constraints in the classroom to compile an appropriate test. We also checked the distribution of correct answers to ensure that there were no systematic patterns (e.g., B, C, D as correct answers to successive questions), nor a preponderance of one correct letter choice over the others.

For all of our validation procedures, including the cognitive interviews and our pilot testing, we obtained human subjects research approval from our institution. It is important to note that both your instrument validation studies and your field research or evaluation studies may be considered human subjects research, depending on the steps you take (National Science Foundation [NSF], 2015; U.S. Department of Health and Human Services, 2015). Therefore, research approval may need to be obtained from your institution’s institutional review board (IRB), the committee that oversees research involving human subjects. Before starting your project, you will need to check with your institution on the requirements for conducting research and publishing the results.

SUMMARY AND CONCLUSIONS

The findings from an evaluation of a K–12 educational intervention or the claims made about student performance in an undergraduate or graduate course depend greatly on the quality of the instruments used to assess learning. In this essay, we outlined an iterative four-step process for developing and validating items for small-scale, low-stakes research and instructional contexts. First, identify the intervention’s or curriculum’s broad learning goals. Then, outline and prioritize the specific learning objectives you wish to measure. Second, seek out instruments that address those objectives and/or draft new items. Consider how well the items fit the knowledge and reading level of your target audience and conform to recommended item-writing guidelines (Haladyyna et al., 2002). Sometimes you may have to modify existing items. Third, as you prepare items, consider the number of points to assign to each answer and establish clear, reliable scoring criteria for open-ended questions. Fourth, validate your instruments using a variety of data sources, including expert review, cognitive interviews, and testing in classrooms. Collectively, these data provide the backing for warranted arguments supporting the instruments’ interpretation and use (DeBarger et al., 2013; Kane, 2013).

In our experience, we have learned three particularly salient lessons about creating measures.

1. Plan your development efforts by clearly describing what you are measuring and the context in which you are collecting your measurements. It can be tempting to rush into writing items without explicitly articulating your instrument targets, but resist this inclination. The more work you can do up front to describe your intended instrument, the more efficient you can be in the later phases of development. In fact, one of the authors (K.B.) often organizes her thoughts on an instrument cover sheet (see the Supplemental Material) before she begins researching or drafting items. As we have also demonstrated throughout this article, the prior knowledge of the individuals taking the assessments can have a significant influence on their interpretation of items. Reminding yourself of this audience up front can help you construct your questions accordingly. Finally, you may want to think about the time and resources available for data analysis, as this can influence your decision to use easily scored MC items or more labor-intensive open-ended or other types of items.

2. Leverage a community of experts to help design and review your assessments. Assessment development is typically not an individual enterprise, but a community affair. We encourage (when possible) collaborations between assessment developers and curriculum designers to prioritize the content to be assessed and to draft items. If needed, consult with content experts to make sure the items accurately represent the ideas the instrument is intended to measure, educators to ensure that the wording of the items is appropriate for the students who will be taking the assessment,
and assessment-takers to identify any unanticipated misinterpretations of test questions. It may also be beneficial to work with psychometricians, external evaluators, or other researchers with expertise in assessment development to add rigor to your work. Some universities have teaching and learning centers with staff members who may be able to help you develop new assessments or improve existing ones. We encourage you to consult these and any other experts you may be able to access.

3. Allocate more time than you expect. Developing assessments is not a one-time, one-sitting process, but requires multiple rounds of planning, drafting, review, and revision. The amount of time required to construct an instrument varies based on a) the complexity and number of constructs to be measured, b) the availability of existing instruments, c) the necessity of securing IRB approval to pilot test items, and d) the ease or difficulty of recruiting students or classes who can participate in pilot testing within your preferred time frame.

We believe that investing in constructing quality measures is essential for advancing biology education research and improving the pedagogy of science faculty. As noted in the Introduction, vetted assessments are critical for identifying evidence-based curricula and practices that improve student learning. Research and evaluation studies that do not use validated measures are limited in the strength of the claims they can make about the efficacy of the materials or program they investigated. As funding agencies increasingly require projects to demonstrate broad societal impact (NSF, 2007), plans for research and evaluation with thoroughly crafted instruments can influence panelists’ decisions to recommend a proposal for funding. Moreover, faculty whose career advancement, retention, or promotion depends in part on demonstrating advancements in education need to provide rigorous evidence of their accomplishments. Data from instruments shown to be valid for their intended purpose can carry significant weight in this regard. Finally, assessment quality is one of the factors considered by manuscript reviewers, since it is part of judging the approach used to study a teaching practice or evaluate an intervention.

If the steps outlined in this paper initially seem daunting, start by asking yourself two questions: “What exactly do I want students to know and do by the end of this course/set of curricular materials?” and “What evidence would convince me that students have learned the core objectives?” Your answers will help you begin to think about the questions you should be asking in your assessments.

You can also take small steps to improve your existing instruments. For example, begin by reviewing an MC test you have used in a recent course. Look at the percentage of students who got each item correct. Were there any items that you thought more students would answer correctly? Were there items that you thought would be difficult and that you expected fewer students to answer correctly? For open-ended questions, you could take a similar tally of scores. If you find that more than 80% or less than 30% of the students got full credit on any test question, those items may be too easy or difficult for you to say anything meaningful about learning.

Pick five to seven items that seem problematic and decide how you will handle them. Is the concept addressed by the question important to assess? If so, modify it. If not, replace it. Use the strategies in this primer to guide your revisions and consider reaching out to colleagues for assistance. Incorporate the revised items into a new assessment (even a comprehensive final exam) and see how students perform. You might even include the old items on the same test for comparison. Are you now more confident about what you can say about what students have learned?

We have found that assessment development requires creativity, collaboration, and persistence, much like many other scientific endeavors. With time and experience, the process can become easier, though it will always involve challenging decisions about constructs, observations, and interpretations. We are confident that these efforts are worthwhile. Engaging in an evidence-based process of assessment item development can go a long way toward improving the contributions from small-scale biology education projects.

ACKNOWLEDGMENTS

We thank Molly Malone, the GSLC’s senior education specialist, who helped us prioritize the constructs we measured in our assessments and reviewed drafts of items; Sheila Homburger, the GSLC’s science content manager, who also reviewed drafts of items; the students who participated in cognitive interviews; the teachers who allowed us to pilot test the instruments in their classrooms; LSE monitoring editor Dr. Ross Nehm and the reviewers who provided valuable feedback on the manuscript; and Dr. George Bass for his advice on a portion of the article. The project described was supported by award number R25RR023288 from the National Center for Research Resources (NCRR). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NCRR or the National Institutes of Health.

REFERENCES


American Association for the Advancement of Science Project 2061 (2011). Developing and using assessments aligned to science learning goals. Workshop held 12–14 October 2011, in Washington, DC.


Bowling B, Acra E, Wang L, Myers MF, Dean GE, Markle GC, Moskalik CL, Huether CA (2008). Development and evaluation...
of a genetics literacy assessment instrument for undergraduates. Genetics 178, 15–22.


College Board (2009). College Board Standards for College Success (Science), New York: College Board.


Most discipline-based education researchers (DBERs) were formally trained in the methods of scientific disciplines such as biology, chemistry, and physics, rather than social science disciplines such as psychology and education. As a result, DBERs may have never taken specific courses in the social science research methodology—either quantitative or qualitative—on which their scholarship often relies so heavily. One particular aspect of (quantitative) social science research that differs markedly from disciplines such as biology and chemistry is the instrumentation used to quantify phenomena. In response, this Research Methods essay offers a contemporary social science perspective on test validity and the validation process. The instructional piece explores the concepts of test validity, the validation process, validity evidence, and key threats to validity. The essay also includes an in-depth example of a validity argument and validation approach for a test of student argument analysis. In addition to DBERs, this essay should benefit practitioners (e.g., lab directors, faculty members) in the development, evaluation, and/or selection of instruments for their work assessing students or evaluating pedagogical innovations.

INTRODUCTION

The field of discipline-based education research (Singer et al., 2012) has emerged in response to long-standing calls to advance the status of U.S. science education at the post-secondary level (e.g., Boyer Commission on Educating Undergraduates in the Research University, 1998; National Research Council, 2003; American Association for the Advancement of Science, 2011). Discipline-based education research applies scientific principles to study postsecondary science education processes and outcomes systematically to improve the scientific enterprise. In particular, this field has made significant progress with respect to the study of active-learning pedagogies (e.g., Freeman et al., 2014); 2) interventions to support those pedagogies among both faculty (e.g., Brownell and Tanner, 2012) and graduate teaching assistants (e.g., Schussler et al., 2015); and 3) undergraduate research experiences (e.g., Auchincloss et al., 2014).

Most discipline-based education researchers (DBERs) were formally trained in the methods of scientific disciplines such as biology, chemistry, and physics, rather than social science disciplines such as psychology and education. As a result, DBERs may have never taken specific courses in the social science research methodology—either quantitative or qualitative—on which their scholarship often relies so heavily (Singer et al., 2012). While the same principles of science ground all these fields, the specific methods used and some criteria for methodological and scientific rigor differ along disciplinary lines.

One particular aspect of (quantitative) social science research that differs markedly from research in disciplines such as biology and chemistry is the instrumentation used to quantify phenomena. Instrumentation is a critical aspect of research methodology, because it provides the raw materials input to statistical analyses and thus serves as a basis for credible conclusions and research-based educational practice (Opfer et al., 2012; Campbell and Nehm, 2013). A notable feature of social science instrumentation is that it generally targets variables that are latent, that is, variables...
that are not directly observable but instead must be inferred through observable behavior (Bollen, 2002). For example, to elicit evidence of cognitive beliefs, which are not observable directly, respondents are asked to report their level of agreement (e.g., “strongly disagree,” “disagree,” “agree,” “strongly agree”) with textually presented statements (e.g., “I like science,” “Science is fun,” and “I look forward to science class”). Even a multiple-choice final examination does not directly observe the phenomenon of interest (e.g., student knowledge). As such, compared with work in traditional scientific disciplines, in the social sciences, more of an inferential leap is often required between the derivation of a score and its intended interpretation (Opfer et al., 2012).

Instruments designed to elicit evidence of variables of interest to DBERs have proliferated in recent years. Some well-known examples include the Experimental Design Ability Test (EDAT; Sirum and Humbrung, 2011); the Genetics Concept Assessment (Smith et al., 2008); the Classroom Undergraduate Research Experience survey (Denefrio et al., 2007); and the Classroom Observation Protocol for Undergraduate STEM (Smith et al., 2013). However, available instruments vary widely in their quality and nuance (Opfer et al., 2012; Singer et al., 2012; Campbell and Nehm, 2013), necessitating understanding on the part of DBERs of how to evaluate instruments for use in their research. Practitioners, too, should know how to evaluate and select high-quality instruments for program evaluation and/or assessment purposes. Where high-quality instruments do not already exist for use in one’s context, which is commonplace (Opfer et al., 2012), they need to be developed, and corresponding empirical validity evidence needs to be gathered in accord with contemporary standards.

In response, this Research Methods essay offers a contemporary social science perspective on test validity and the validation process. It is intended to offer a primer for DBERs who may not have received formal training on the subject. Using examples from discipline-based education research, the instructional piece explores the concepts of test validity, the validation process, validity evidence, and key threats to validity. The essay also includes an in-depth example of a validity argument and validation approach for a test of student argument analysis. In addition to DBERs, this essay should benefit practitioners (e.g., lab directors, faculty members) in the development, evaluation, and/or selection of instruments for their work assessing students or evaluating pedagogical innovations.

**TEST VALIDITY AND THE TEST VALIDATION PROCESS**

A test is a sample of behavior gathered in order to draw an inference about some domain or construct within a particular population (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education [AERA, APA, and NCME], 2014). In the social sciences, the domain about which an inference is desired is typically a latent (unobservable) variable. For example, the STEM GTA-Teaching Self-Efficacy Scale (DeChenne et al., 2012) was developed to support inferences about the degree to which a graduate teaching assistant believes he or she is capable of 1) cultivating an effective learning environment and 2) implementing particular instructional strategies. As another example, the inference drawn from an introductory biology final exam is typically about the degree to which a student understands content covered over some extensive unit of instruction. While beliefs or conceptual knowledge are not directly accessible, what can be observed is the sample of behavior the test elicits, such as test-taker responses to questions or responses to rating scales. Diverse forms of instrumentation are used in discipline-based education research (Singer et al., 2012). Notable subcategories of instruments include self-report (e.g., attitudinal and belief scales) and more objective measures (e.g., concept inventories, standardized observation protocols, and final exams). By the definition of “test” above, any of these instrument types can be conceived as tests—though the focus here is only on instruments that yield quantitative data, that is, scores.

The paramount consideration in the evaluation of any test’s quality is validity: “the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (Angoff, 1988; AERA, APA, and NCME, 2014, p. 11). In evaluating test validity, the focus is not on the test itself, but rather the proposed inferences drawn on the basis of the test’s score(s). Noteworthy in the validity definition above is that validity is a matter of degree (“the inferences supported by this test have a high or low degree of validity”), rather than a dichotomous character (e.g., “the inferences supported by this test are or are not valid”).

Assessment validation is theorized as an iterative process in which the test developer constructs an evidence-based argument for the intended test-based score interpretations in a particular population (Kane, 1992; Messick, 1995). An example validity argument claim is that the test’s content (e.g., questions, items) is representative of the domain targeted by the test (e.g., body of knowledge/skills). With this argument-based approach, claims within the validity argument are substantiated with various forms of relevant evidence. Altogether, the goal of test validation is to accumulate over time a comprehensive body of relevant evidence to support each intended score interpretation within a particular population (i.e., whether the scores should in fact be interpreted to mean what the developer intends them to mean).

**CATEGORIES OF TEST VALIDITY EVIDENCE**

Historically, test validity theory in the social sciences recognized several categorically different “types” of validity (e.g., “content validity,” “criterion validity”). However, contemporary validity theory posits that test validity is a unitary (single) concept. Rather than providing evidence of each “type” of validity, the charge for test developers

---

1 A test cannot be “stamped” valid for all purposes and test-taker populations; validity evidence needs to be gathered with respect to all intended instrument uses.

2 While other key dimensions for evaluating an instrument’s quality include reliability (i.e., test score consistency) and utility (i.e., feasibility; AERA, APA, and NCME, 2014), the focus here is on validity only.

3 While this essay allies with test validity theory as codified in the Standards for Educational and Psychological Testing (AERA, APA, and NCME, 2014), the reader will note that there are alternative conceptions of validity as well (Lissitz and Samuelson, 2007).
is to construct a cohesive argument for the validity of test score–based inferences that integrates different forms of validity evidence. The categories of validity evidence include evidence based on test content, evidence based on response processes, evidence based on internal structure, evidence based on relations with other variables, and evidence based on the consequences of testing (AERA, APA, and NCME, 2014). Figure 1 provides a graphical representation of the categories and subcategories of validity evidence.

Validity evidence based on test content concerns “the relationship between the content of a test and the construct it is intended to measure” (AERA, APA, and NCME, 2014, p. 14). Such validity evidence concerns the match between the domain purportedly measured by (e.g., diagnostic microscopy skills) and the content of the test (e.g., the specific slides examined by the test taker). For example, if a test is intended to elicit evidence of students’ understanding of the key principles of evolution by means of natural selection (e.g., variation, heredity, differential fitness), the test should fully represent those principles in the sample of behavior it elicits. As a concrete example from the literature, in the development of the Host-Pathogen Interaction (HPI) concept inventory, Marbach-Ad et al. (2009) explicitly mapped each test item to one of 13 HPI concepts intended to be assessed by their instrument. Content validity evidence alone is insufficient for establishing a high degree of validity; it should be combined with other forms of evidence to yield a strong evidence-based validity argument marked by relevancy, accuracy, and sufficiency.

In practice, providing validity evidence based on test content involves evaluating and documenting content representativeness. One standard approach to collecting evidence of content representativeness is to submit the test to external systematic review by subject matter–area experts (e.g., biology faculty) and to document such reviews (as well as revisions made on their basis). External reviews focus on the adequacy of the test’s overall elicited sample of behavior in representing the domain assessed and any corresponding subdomains, as well as the relevance or irrelevance of particular questions/items to the domain. We refer the reader to Webb (2006) for a comprehensive and sophisticated framework for evaluating different dimensions of domain–test content alignment.

Another approach used to design a test, so as to support and document construct representativeness, is to employ a “table of specifications” (e.g., Fives and DiDonato-Barnes, 2013). A table of specifications (or test blueprint) is a tool for designing a test that classifies test content along two dimensions, a content dimension and a cognitive dimension. The content dimension pertains to the different aspects of the construct one intends to measure. In a classroom setting, aspects of the construct are typically defined by behavioral/instructional objectives (i.e., students will analyze phylogenetic trees). The cognitive dimension represents the level of cognitive processing or thinking called for by test components (e.g., knowledge, comprehension, analysis). Within a table of specifications, one indicates the number/percent of test questions or items for each aspect of the construct at each
cognitive level. Often, one also provides a summary measure of the number of items pertaining to each content area (regardless of cognitive demand) and at each cognitive level (regardless of content). Instead of or in addition to the number of items, one can also indicate the number/percent of available points for each content area and cognitive level. Because a table of specifications indicates how test components represent the construct one intends to measure, it serves as one source of validity evidence based on test content. Table 1 presents an example table of specifications for a test concerning the principles of evolution by means of natural selection.

Evidence of validity based on response processes concerns “the fit between the construct and the detailed nature of the performance or response actually engaged in by test takers” (AERA, APA, and NCME, 2014, p. 15). For example, if a test purportedly elicits evidence of undergraduate students’ critical evaluative thinking concerning evidence-based scientific arguments, during the test the student should be engaged in the cognitive process of examining argument claims, evidence, and warrants, and the relevance, accuracy, and sufficiency of evidence. Most often one gathers such evidence through respondent think-aloud procedures. During think alouds, respondents verbally explain and rationalize their thought processes and responses concurrently during test completion. One particular method commonly used by professional test vendors to gather response process–based validity evidence is cognitive labs, which involve both con- temporary and retrospective verbal reporting by respondents (Willis, 1999; Zucker et al., 2004). As an example from the literature, developers of the HPI concept inventory asked respondents to provide open-ended responses to ensure that their reasons for selecting a particular response option (e.g., “B”) were consistent with the developer’s intentions, that is, the student indeed held the particular alternative conception presented in response option B (Marbach-Ad et al., 2009). Think alouds are formalized via structured protocols, and the elicited think-aloud data are recorded, transcribed, analyzed, and interpreted to shed light on validity.

Evidence based on internal structure concerns “the degree to which the relationships among test item and test components conform to the construct on which the proposed test score interpretations are based” (AERA, APA, and NCME, 2014, p. 16). For instance, suppose a professor plans to teach one topic (eukaryotes) using small-group active-learning instruction and another topic (prokaryotes) through lecture instruction; and he or she wants to make within-class comparisons of the effectiveness of these methods. As an outcome measure, a test may be designed to support inferences about the two specific aspects of biology content (e.g., characteristics of prokaryotic and eukaryotic cells). Collection of evidence based on internal structure seeks to confirm empirically whether the scores reflect the (in this case two) distinct domains targeted by the test (Messick, 1995). In practice, one can formally establish the fidelity of test scores to their theorized internal structure through methodological techniques such as factor analysis, item response theory, and Rasch modeling (Harman, 1960; Rasch, 1960; Embretson and Reise, 2013). With factor analysis, for example, item intercorrelations are analyzed to determine whether particular item responses cluster together (i.e., whether scores from components of the test related to one aspect of the domain [e.g., questions about prokaryotes] are more interrelated with one another than they are with scores derived from other components of the test [e.g., questions about eukaryotes]).

Item response theory and Rasch models hypothesize that the probability of a particular response to a test item is a function of the respondent’s ability (in terms of what is being measured) and characteristics of the item (e.g., difficulty, discrimination, pseudo-guessing). Examining test score internal structure with such models involves examining whether such model-based predictions bear out in the observed data. There are a variety of such models for use with test questions with different (or different combinations of) response formats such as the Rasch rating-scale model (Andrich, 1978) and the Rasch partial-credit Rasch model (Wright and Masters, 1982).

Validity evidence based on relations with other variables concerns “the relationship of test scores to variables external to the test” (AERA, APA, and NCME, 2014, p. 16). The collection of this form of validity evidence centers on examining how test scores are related to both measures of the same or similar constructs and measures of distinct and different constructs (i.e., respectively termed “convergent validity” and “discriminant validity” evidence). In other words, such evidence pertains to how scores relate to other variables as would be theoretically expected. For example,

Table 1. Example table of specifications for evolution by means of natural selection test showing numbers of test items pertaining to each content area at each cognitive level and total number of items per content area and cognitive level

<table>
<thead>
<tr>
<th>Content (behavioral objective)</th>
<th>Comprehension</th>
<th>Application</th>
<th>Analysis</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Students will define evolution by means of natural selection.</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>2. Students will define key principles of evolution by means of natural selection (e.g., heredity, differential fitness).</td>
<td>5</td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>3. Students will compute measures of absolute and relative fitness.</td>
<td>5</td>
<td>3</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>4. Students will compare evolution by means of natural selection with earlier evolution theories.</td>
<td>4</td>
<td>4</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>5. Student will analyze phylogenetic trees.</td>
<td>6</td>
<td>5</td>
<td>7</td>
<td>18</td>
</tr>
</tbody>
</table>

This source of evidence has been termed “substantive validity” (Messick, 1995).
if a new self-report instrument purports to measure experimental design skills, scores should correlate highly with an existing measure of experimental design ability such as the EDAT (Sirum and Humburg, 2011). On the other hand, scores derived from this self-report instrument should be considerably less correlated or uncorrelated with scores from a personality measure such as the Minnesota Multiphasic Personality Inventory (Greene, 2000). As another discipline-based education research example, Nehm and Schonfeld (2008) collected discriminant validity evidence by relating scores from both the Conceptual Inventory of Natural Section (CINS) and the Open Response Instrument (ORI), which both purport to assess understanding of and conceptions concerning natural selection, and a geology test of knowledge about rocks.

A subcategory of evidence based on relations with other variables is evidence related to test-criterion relationships, which concerns how test scores are related to some other non-test indicator or outcome either at the same time (so-called concurrent validity evidence) or in the future (so-called predictive validity evidence). For instance, developers of a new biostatistics test might examine how scores from the test correlate as expected with professor ability judgments or mathematics course grade point average at the same point in time; alternatively, the developer might follow tested individuals over time to examine how scores relate to the probability of successfully completing biostatistics course work. As another example, given prior research on self-efficacy, scores from instruments that probe teaching self-efficacy should correlate highly with scores derived from this self-report instrument should be considerably less correlated or uncorrelated with scores from other examples, given prior research on self-efficacy, scores from instruments that probe teaching self-efficacy should be related to respondents’ levels of teacher training and experience (Prieto and Altmair, 1994; Prieto and Meyers, 1999).

Examination of how test scores are related or not to other variables as expected is often associational in nature (e.g., correlational analysis). There are also two other specific methods for eliciting such validity evidence. The first is to examine score differences between theoretically different groups (e.g., whether scientists’ and nonscientists’ scores from an experimental design test differ systematically on average)—the “known groups method.” The second is to examine whether scores increase or decrease as expected in response to an intervention (Hattie and Cooksey, 1984; AERA, APA, and NCME, 2014). For example, Marbach-Ad et al. (2009, 2010) examined HFI concept inventory score differences between majors and nonmajors and students in introductory and upper-level courses. To inform the collection of validity evidence based on relations with other variables, individuals should consult the literature to formulate a theory around how good measures of the construct should relate to different variables. One should also note that the quality of such validity evidence hinges on the quality (e.g., validity) of measures of external variables.

Finally, validity evidence based on the consequences of testing concerns the “soundingness of proposed interpretations of test scores” for their intended uses” (AERA, APA, and NCME, 2014, p. 19) and the value implications and social consequences of testing (Messick, 1994, 1995). Such evidence pertains to both the intended and unintended effects of test score interpretation and use (Linn, 1991; Messick, 1995).

Example intended consequences of test use would include motivating students, better-targeted instruction, and populating a special program with only those students who are in need of the program (if those are the intended purposes of test use). An example of an unintended consequence of test use would be significant reduction in instructional time because of overly time-consuming test administration (assuming, of course, that this would not be a desired outcome) or drop out of particular student populations because of an excessively difficult test administered early in a course. In K–12 settings, a classic example of an unintended consequence of testing is the “narrowing of the curriculum” that occurred as a result of the No Child Left Behind Act testing regime; when faced with annual tests focused only on particular content areas (i.e., English/language arts and mathematics), schools and teachers focused more on tested content and less on non-tested content such as science, social studies, art, and music (e.g., Berliner, 2011). Evidence based on the consequences of a test is often gathered via surveys, interviews, and focus groups administered with test users.

**TEST VALIDITY ARGUMENT EXAMPLE**

In this section, we provide an example validity argument for a test designed to elicit evidence of students’ skills in analyzing the elements of evidence-based scientific arguments. This hypothetical test presents text-based arguments concerning scientific topics (e.g., global climate change, natural selection) to students, who then directly interact with the texts to identify their elements (i.e., claims, reasons, and warrants). The test is intended to support inferences about 1) students’ overall evidence-based science argument-element analysis skills; 2) students’ skills in identifying particular evidence-based science argument elements (e.g., claims); and 3) errors made when students identify particular argument elements (e.g., evidence). Additionally, the test is intended to 4) support instructional decision-making to improve science teaching and learning. The validity argument claims undergirding this example test’s score interpretations and uses (and the categories of validity evidence advanced to substantiate each) are shown in Table 2.

**ANALYSIS OF CINS VALIDITY EVIDENCE**

The example validity argument provided in the preceding section was intended to model the validity argument formulation process for readers who intend to develop an instrument. However, in many cases, an existing instrument (or one of several existing instruments) needs to be selected for use in one’s context. The use of an existing instrument for research or practice requires thoughtful analysis of extant validity evidence available for an instrument’s score interpretations and uses. Therefore, in this section, we use validity theory as outlined in the Standards for Educational and Psychological Testing to analyze the validity evidence for a particular instrument, the CINS.

As reported in Anderson et al. (2002), the CINS is purported to measure “conceptual understanding of natural selection” (as well as alternative conceptions of particular relevant ideas diagnostically) in undergraduate non-biology...
The two primary threats to test score validity are construct underrepresentation and construct-irrelevant variance. Construct underrepresentation is “the degree to which a test fails to capture important aspects of the construct.”
(AERA, APA, and NCME, 2014; p. 12). In other words, construct underrepresentation involves failing to elicit a representative sample of behavior from test takers (e.g., responses to multiple-choice questions) relative to the universe of possible relevant behaviors that might be observed. While it is neither necessary nor feasible to ask respondents to engage in every single possible relevant behavior, it is crucial that the behavior sampled by the test is sufficiently representative of the construct at large. If a test does not fully and adequately sample behavior related to the targeted domain, the test score’s meaning in actuality would be narrower than is intended.

Content underrepresentation can be mitigated by initiating test design with a thorough analysis and conception of the domain targeted by the test (Mislevy et al., 2003; Opfer et al., 2012). Knowledge of the construct, and variables that are related or not related to the construct, can also inform the validation process (Mislevy et al., 2003). Beginning test design with a thorough conception of the construct one intends to measure is analogous to the course design approach known as “backward design”; with backward design one first identifies what one wants students to know/be able to do after instruction (learning outcomes) and then designs a course to get students there (Wiggins and McTighe, 2005). Other strategies to promote construct representation include building a test based on a table of specifications; submitting a text to external expert content review (as both noted above); and employing a sufficient number of test items to ensure good representation of domain content.

Besides construct representation, the other primary threat to test score validity is construct-irrelevant variance—“the degree to which test scores are affected by processes that are extraneous to the test’s intended purpose” (AERA, APA, and NCME, 2014, p. 12). Construct-irrelevant variance is test score variation caused systematically by factors other than (or in addition to) those intended; in other words, some part of the reason why one received a “high” or “low” score is due to irrelevant reasons. Two common examples of this are: English skills affecting test scores for non-native English speakers on tests written in English; and computer skills affecting test scores for tests administered via computer. Another example would be if items on a science teaching self-efficacy self-report instrument are written so generally that the scores represent not science teaching-specific self-efficacy but self-efficacy in general. It is critical to mitigate such threats through test design processes (e.g., minimizing test linguistic load). One can often identify potential threats in the course of a thorough analysis of the construct/domain done at early design stages. During test validation one should also disconfirm such threats wherein scores are driven by irrelevant factors; practitioners often conduct factor, correlational, and differential item functioning analyses toward this end.

**SUMMARY**

Systematic research on postsecondary science teaching and learning and evaluation of local innovations by practitioners hinges on the availability and use of sound instrumentation. Unfortunately, the field of discipline-based education research lacks sufficient existing and high-quality instruments for use in all of these efforts (Opfer et al., 2012; Singer et al., 2012; Campbell and Nehm, 2013). DBERs and practitioners furthermore do not typically have formal training that equips them to evaluate and select existing instruments or develop and validate their own instruments for needed purposes. This essay reviewed contemporary test validity and validation theory for DBERs and practitioners in hopes of equipping them with such knowledge.

This essay was chiefly intended for two audiences: 1) those who will develop new instruments; and 2) those who will evaluate and select from among existing instruments. Here, we summarize the implications of this essay for members of these two populations. First, it behooves those developing and publishing their own instruments to explicitly frame, construct, and report an evidence-based validity argument for their proposed instruments’ intended score interpretations and uses. This argument should rely on multiple forms of validity evidence and specify the test-taker and user populations for which that argument pertains. If faced with space constraints in journal articles, test manuals or technical reports can be written to detail such validity evidence and made available to the scholarly community.

Like any argument, an evidence-based argument formulated during test validation should be characterized by relevance, accuracy, and sufficiency. As such, validity arguments should be held up to scientific scrutiny before a test’s operational use. The quality of a validity argument hinges on a number of factors discussed in this essay. Examples include the alignment of the validity argument claims with intended test score interpretations and uses; the representativeness of the samples from which validity evidence is gathered to the intended test-taker population; the relevance of the expertise held by content reviewers; and the technical quality of external measures. A final point to emphasize is that validation is an ongoing process; additional validity evidence may need to be gathered as theory concerning a construct evolves or as the community seeks to use an instrument with new populations.

Second, potential test users should be critical in their evaluation of existing instruments, and should not merely assume a strong validity argument exists for an instrument’s score interpretations and uses with a particular population. Potential users should look to the instrumentation (or methods) sections of published articles for key information, such as whether the test was developed based on a sound theoretical conception of construct, whether the test underwent external content review, and whether scores correlate with other measures as they theoretically should, among other things. One may have to contact an author for such information. Altogether, such practices should advance the quality of measurement within the realm of discipline-based education research.

---

4While our focus is on instruments comprising sets of questions or items intended to elicit evidence of a particular construct or constructs, many of the ideas here apply also to questionnaire (survey) validation. For example, the developer of a questionnaire may interrogate how respondents interpret and formulate a response to a particular question as validity evidence based on response processes.
ACKNOWLEDGMENTS

The authors thank Drs. Beth Schussler and Ross Nehm and two anonymous reviewers for their constructive feedback on an earlier version of this article.

REFERENCES

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


Wright BD, Masters GN (1982). Rating Scale Analysis, Chicago, IL: MESA.

Recent biomedical workforce policy efforts have centered on enhancing career preparation for trainees, and increasing diversity in the research workforce. Postdoctoral scientists, or postdocs, are among those most directly impacted by such initiatives, yet their career development remains understudied. This study reports results from a 2012 national survey of 1002 American biomedical postdocs. On average, postdocs reported increased knowledge about career options but lower clarity about their career goals relative to PhD entry. The majority of postdocs were offered structured career development at their postdoctoral institutions, but less than one-third received this from their graduate departments. Postdocs from all social backgrounds reported significant declines in interest in faculty careers at research-intensive universities and increased interest in nonresearch careers; however, there were differences in the magnitude and period of training during which these changes occurred across gender and race/ethnicity. Group differences in interest in faculty careers were explained by career interest differences formed during graduate school but not by differences in research productivity, research self-efficacy, or advisor relationships. These findings point to the need for enhanced career development earlier in the training process, and interventions sensitive to distinctive patterns of interest development across social identity groups.
Researchers have also established strong relationships between self-efficacy and retention and success in science, particularly for younger trainees. For example, self-efficacy is one of the core constructs of social cognitive career theory (SCCT; Lent et al., 1994), suggesting that individuals’ confidence in their ability to complete tasks related to their fields of interest is related to the formation of career interests and persistence in educational and occupational pursuits. Self-efficacy may serve as the foundation for developing outcome expectations, career interests, and career goals in a given field (Lent et al., 2008). Science self-efficacy mediated the relationship between support services and commitment to science careers for graduate and undergraduate students of color (Chemers et al., 2011), and women who persisted in science, technology, engineering, and mathematics careers demonstrated high rates of academic and relational self-efficacy (Zelden and Pajares, 2000). Further, identifying as a scientist and sense of belonging appear central to this process, particularly for women and men of color. Sense of belonging has long been connected to student retention and success in undergraduate education (Hurtado and Carter, 1997; Hausmann et al., 2007). Again, it remains unclear whether or how these factors influence career development in postdocs.

The goal of this work is to begin to fill some of these knowledge gaps by illuminating the career development of American biomedical postdoctoral scientists. Further, it aims to integrate the sometimes distinct conversations around biomedical career development generally and enhancing gender and racial/ethnic diversity within the research workforce. Specifically, this work aims to:

1. describe the career development of postdocs, examining:
   a. how career interests, knowledge, and career goal clarity evolve over the course of training;
   b. perceptions of access to structured career development during training;
   c. perceptions of career support received from advisors and institutions;
2. understand the extent to which determinants previously linked to career attainment (e.g., research self-efficacy, productivity, training experiences) predict interest in various career pathways; and
3. examine the extent to which these patterns are overlapping and distinct by social identity (race/ethnicity, gender, and their intersection).

METHODS

Data Collection and Procedures

All work was done under the approval of the University of Maryland Institutional Review Board (IRB #373799-5), and all respondents provided consent for participation in the study. The survey instrument and survey data-collection procedures have been previously published (Gibbs et al., 2014). Participants completed a short survey that focused on their graduate and postdoctoral training experiences, career development, and professional interests. There were a total of 1890 complete and unique responses. This analysis focuses on the respondents (n = 1002) who met the following criteria: 1) self-identified as a U.S. citizen or

(Continued on next page)
permanent resident, 2) held a postdoctoral position at the time of survey completion (October 2012-January 2013), and 3) completed a PhD in the biomedical or behavioral sciences (as defined by the NIH Biomedical Research Workforce Working Group Report; NIH, 2012) between 2007 and 2012. Information on the participants’ disciplinary backgrounds, PhD, and postdoctoral training institutions is available in Supplemental Material Tables S1–S3.

Although noncitizen postdocs are the majority of all postdocs in the United States, analysis of career development and interest was restricted to citizens and permanent residents because of issues (such as visa status) that uniquely shape their career decision making. The sampling strategy does not permit the calculation of a formal response rate; however, the sample represents ~8.9% of eligible respondents (i.e., U.S. citizen and permanent resident postdocs in the life sciences, bioengineering, neuroscience, and psychology [henceforth, “biomedical sciences”] in 2012) and 18.8% of eligible postdocs from URM backgrounds in this category (National Science Foundation, 2013).

Social Identity. Social identity was stratified based on the intersections of race/ethnicity and gender as previously described (Gibbs et al., 2014). Males from well-represented racial/ethnic backgrounds (i.e., white, Asian/Asian American, or both white and Asian/Asian American) are referred to as WRM (25.6% of sample; n = 257); males from URM racial/ethnic (American Indian/Alaska Native, black/African-American, Hispanic/Latino, or Native Hawaiian/Pacific Islander) backgrounds are referred to as URMM (5.4% of sample; n = 54); females from well-represented backgrounds are referred to as WRF (54.6% of sample; n = 547); and females from URM backgrounds are referred to as URMF (12.2% of sample; n = 122).

Career Interest Measures. Respondents reported their interest in four career pathways at three time points: 1) the beginning of their PhD training, 2) the completion of their PhD training, and 3) currently. These pathways were:

- Faculty at a research-intensive university
- Faculty at a teaching-intensive university
- Research career, nonacademic (e.g., industry, pharmaceutical, biotech, government, start-up)
- Nonresearch career (e.g., consulting, policy, science writing, patent law, business)

Interest was measured on a six-point scale, on which 0 represented not knowledgeable; 1, no interest; 3, moderate interest; and 5, strong interest. For this analysis, respondents answering not knowledgeable were recoded as having no interest. Sensitivity analysis revealed the inferences about career development did not change based on this reclassification. Those who answered 4 or 5 were considered “high interest” (see Figure 2 and Tables 1 and 2 later in this article).

Career Development, Training Experience, and Research Self-Efficacy Measures. Respondents rated their level of agreement with statements about their career goal clarity and knowledge of career options available to PhDs in their disciplines at three time points: 1) the beginning of their PhD training, 2) the completion of their PhD training, and 3) currently.

Respondents also rated their level of agreement with statements regarding their graduate and postdoctoral training experiences. Measures focused on self-reported:

- Sense of belonging, both intellectually and socially: This was assessed in respondents’ 1) graduate research groups, 2) graduate departments, and 3) postdoctoral research groups.
- Advisor interactions: The extent to which respondents felt their graduate and postdoctoral advisors were 1) “invested” in their career advancement and 2) “equally supportive of [trainees] pursuing academic and non-academic career paths.”
- Career development: The extent to which respondents 1) were “offered structured opportunities to explore a variety of career pathways (academic and non-academic)” by their graduate departments and postdoctoral institutions and 2) felt their graduate departments were “equally supportive of [trainees] pursuing academic and non-academic career paths.”
- Research self-efficacy: This was assessed via the respondent’s agreement with the statement “I am confident in my abilities as an independent researcher.”

Agreement was measured on a five-point scale, on which 1 represented strongly disagree; 2, disagree; 3, neither agree nor disagree; 4, agree; and 5, strongly agree. For analytical purposes, those who answered 1 or 2 were considered “disagree,” those who answered 3 were considered “neutral,” and those who answered 4 or 5 were considered “agree” (see Figures 1 and 3 later in this article).

Statistical Analysis. Paired t tests and repeated measures analysis of variance (ANOVA) were used to assess changes in levels of agreement across time points in the overall sample (Figure 1, A and B) and levels of interest within each social identity group (Figure 2) (Fagerland, 2012). The inferences provided by each method of analysis were identical. For Figure 1, A and B, level of agreement at PhD completion or currently (postdoc) was compared with the level of interest at PhD entry (pre–post tests). In Figure 2, within each social identity group, three comparisons were made: 1) PhD entry versus PhD completion (to examine changes during the course of graduate training), 2) PhD completion versus currently (to examine changes during the course of postdoctoral training), and 3) PhD entry versus currently (to examine changes from the beginning of graduate training to postdoctoral training). A Bonferroni-corrected ANOVA was used to compare differences in levels of agreement or career pathway interest between social identity groups at any time point.

Multiple logistic regression analysis was used to determine how career pathway interest was influenced by measures of the career development, training experiences, and research self-efficacy measures (for full details, see Gibbs et al., 2014). Additionally, objective measures (first-author publication rate [i.e., first-authored publications/total years in graduate and postdoctoral training] and total years in postdoctoral training) were also included in the models. Career interest measures at all time points (PhD entry, PhD completion, and currently) were dichotomized into high interest (i.e., 4–5), and low interest (1–3). SEs
Limitations. There are a number of limitations to this work. First, the sample is nonrandom, thus the generalizability of the findings may be limited. While the sample of postdocs from URMM backgrounds is large, the sample of URMM is sometimes underpowered and not able to adequately capture potential differences between their responses and those of postdocs from other groups. The survey did not ask...
about family status (i.e., marriage/partnership, children, and plans for children), which has previously been linked to career attainment in science trainees (Goulden et al., 2009). Additionally, this work relies on self-reported measures of career interest and training experiences, and respondents may have attempted to provide answers that are socially acceptable. We addressed this as previously described (Gibbs et al., 2014). Finally, respondents were asked, in some cases, to retrospectively assess their career interests and training experiences, introducing the potential for incomplete/inaccurate recollections or recall bias. However, understanding respondents’ perceptions of their experiences is important as these are linked to measurable educational and vocational outcomes important for workforce policy such as persistence, goal setting, academic achievement, and satisfaction (Hurtado et al., 1999; Lizzio et al., 2002; Eva et al., 2010).

Figure 2. Changes in career interest patterns across the course of training, by social identity. Postdocs were asked to rate their level interest on a five-point scale (1, no interest; 5, strong interest) in (A) a faculty position at a research-intensive university, (B) a faculty position at a teaching intensive university, (C) a research career outside academia (e.g., industry, pharmaceutical, biotech, government, start-up), and (D) a non-research career (consulting, policy, science writing, patent law, business, etc.) at PhD entry, PhD completion, and currently. Line graphs show the percentage of respondents from each social group reporting high levels of interest (i.e., 4 or 5). The percentage change in each training segment (i.e., from PhD entry to PhD completion; from PhD completion to postdoc) and the overall change (i.e., from PhD entry to currently) are shown. Statistical significance for changes in interest for each group and during each training segment were determined using paired t tests and repeated measures ANOVA. Full statistical analysis is shown in Supplemental Material Table S8.

Level of significance (within group change over time period): **p<0.001  *p<0.05  ^p<0.10
RESULTS

Postdocs Reported Lower Career Goal Clarity and Enhanced Knowledge of Career Options Relative to PhD Entry

Our previous work indicated that PhD scientists often pursued graduate and postdoctoral training without clear career goals and without full knowledge about their potential career options (Gibbs and Griffin, 2013). For this study, respondents were asked to rate their level of agreement at PhD entry, PhD completion, and currently (i.e., during their postdoc) with the following statements: “I had/have a clear career goal,” and “I was/am knowledgeable about the career options available to a person with a PhD in my discipline.” The percentages of postdocs indicating agreement (i.e., a 4 or 5) are shown in Figure 1, A and B.

Overall, at PhD entry, 65% of respondents indicated they had a clear career goal; however, only 26% of postdocs reported having knowledge about the career options available to a person with a PhD in their discipline when they began their PhD training (Figure 1, A and B). A cross-tabulation of these two questions (Supplemental Material Table S4) shows that the majority of those who indicated having a clear career goal at PhD entry (n = 321/627 or 51%) also reported poor knowledge of career options, and only 19% of respondents felt they entered their PhD programs with both a clear career goal and knowledge of career options. As training progressed, greater percentages of postdocs reported knowledge of potential career options (62% at PhD completion and 77% currently; p < 0.001 when comparing knowledge at PhD completion and currently to PhD entry; Figure 1B), yet fewer postdocs reported having clarity about their career goals (54% agreed that they had a clear career goal both at PhD completion and currently; p < 0.001 when comparing clarity at PhD completion and currently with PhD entry). Of note, even though overall career goal clarity decreased, the percentage of those reporting both career goal clarity and knowledge about career options increased to 47% (Supplemental Material Table S5). Collectively, these results suggest early perceptions of career goal clarity occur alongside little knowledge about potential career options available to biomedical PhDs. Further, despite having on average 8 yr of postbaccalaureate training (6 yr PhD, 2 yr postdoctoral training) and enhanced knowledge about available careers, fewer postdocs reported having high levels of career goal clarity than when they started their PhDs. Full data and statistical analysis underlying Figure 1, A and B, are available in Supplemental Material Table S6.

To understand the broader environment in which their career development occurred, postdocs were asked about 1) opportunities for structured career exploration during their graduate and postdoctoral training, 2) the departmental and institutional support offered for pursuit of the full range of career options available to biomedical PhDs, and 3) advisor investment in their career advancement (Figure 1, C–E). Fewer than one-third of respondents (27%) reported that they were offered “structured opportunities to explore a variety of career pathways (academic and non-academic)” by their graduate departments, while nearly two-thirds (65%) reported being offered these structured opportunities for career exploration by their postdoctoral institutions (Figure 1C). Thus, while postdocs reported significant gains in knowledge about career options during graduate school (Figure 1B), these responses suggest that their learning occurred largely outside graduate departments and programs.

Respondents were also asked whether graduate advisors, graduate departments, or postdoctoral advisors were “equally supportive” of students or postdocs “pursuing academic and non-academic career paths” (Figure 1D). Half of the respondents reported their graduate advisors supported all career paths (50%), while fewer than one-third (31%) agreed that their graduate departments were equally supportive of students pursuing all career paths, and slightly more than half (55%) reported their postdoctoral advisors supported postdocs pursuing all career paths. In contrast, three-quarters (75%) of respondents reported that their graduate and postdoctoral advisors were “invested in [their] career advancement” (Figure 1E). Collectively, these results indicate that biomedical PhDs perceive their advisors as supportive of their career development generally but that there is less support for the pursuit of a broad range of career pathways from advisors and departments. Of note, the levels of career goal clarity, knowledge, advisor investment, support for pursuit of all career pathways, and structured career development were largely consistent across social identity (i.e., race/ethnicity and gender; see Supplemental Material Table S7), suggesting that these factors are largely shared for graduate students and postdocs from all backgrounds.

Overall Changes in Career Pathway Interests across Social Identity

Postdocs described their level of interest in four career pathways across the course of their training: 1) faculty at a research-intensive university, 2) faculty at a teaching-intensive university, 3) a research career outside academia, (e.g., industry, pharmaceutical, biotech, government, or a start-up), or 4) a nonresearch career, (e.g., consulting, policy, science writing, patent law, or business). The percentage of postdocs who reported high interest (i.e., a 4 or 5) at PhD entry, PhD completion, and currently is shown in Figure 2 (full data and statistical analysis underlying this figure are available in Supplemental Material Table S8).

Trends in career interests were largely shared across lines of social identity. For all groups, the percentage of postdocs reporting high interest in faculty careers at research universities declined significantly over the course of their training (−10–19% from PhD entry to postdoc, p < 0.02 when comparing levels of interest across time within each social group; Figure 2A); the percentage reporting high interest in faculty careers at teaching universities was unchanged (Figure 2B); the percentage reporting high interest in research careers outside academia increased significantly for WRM, WRW, and URM (+11–17%, p < 0.02 for within-group changes across time; Figure 2C); and the percentage reporting high interest in careers outside research increased significantly (+22–33%; p < 0.0002 for within-group changes; Figure 2D).

There were some notable differences across social identity with respect to both the magnitude and period of training in which changes in interest occurred. During graduate training, there were significant declines in the percentage of women and URM men who expressed high levels of interest in faculty careers at research universities (URRM: −13%,
p = 0.051; WRF: −9%, p < 0.0001; URMF: −12%, p < 0.02); however, there were not significant changes for WRM (−4%; p = 0.20). During postdoctoral training, there were significant declines in the percentage of scientists from well-represented backgrounds reporting high interest in research faculty careers (WRM: −6%, p = 0.01; WRF: −4%; p = 0.008) but not for postdocs from URM backgrounds (URMM: −6%, p = 0.3; URMF: 0%, p = 1.0; Figure 2A). Thus, with respect to interest in faculty careers at research-intensive universities, graduate training appears to be a particularly important time for the development of career interests for women and URM men, while well-represented men may, on average, be engaging in more career decision making during their postdoctoral training. Similar patterns are observed for interest in careers outside research. The increase in the percentage of URMF reporting high interest between PhD entry and completion was greater than that for other groups (+30% for URMF vs. +16–19% for other groups, p < 0.004 when comparing the pre–post change of URMF with all other groups; Figure 2D).

Different Senses of Social "Belonging" and Research Self-Efficacy across Social Identity

Participants’ graduate and postdoc training experiences were also assessed to understand potential factors influencing differences in career interest patterns. We focused on questions related to sense of belonging and research self-efficacy. Postdocs were asked the extent to which they felt they belonged intellectually and socially to their 1) graduate school research group, 2) graduate school department, and 3) postdoctoral research group (Figure 2A, i and B). There were no statistically significant differences across groups in the percentage of postdocs who reported belonging intellectually to their graduate research group (average 87% agreement; Figure 3A, i), belonging intellectually to their postdoctoral research group (average 84% agreement; Figure 3A, iii), or belonging socially to their postdoctoral research group (average 63% agreement; Figure 3B, iii). There were, however, statistically significant differences across social groups in the levels of reported intellectual belonging in their graduate department (Figure 3A, ii; p < 0.03), social belonging in their graduate research group (Figure 3B, i; p < 0.001), and social belonging in their graduate departments (Figure 3B, ii; p = 0.02). In each case, women from URM backgrounds reported the lowest levels of belonging. With respect to social belonging in graduate school, the percentage of URMF who reported belonging was 13–21% lower than their peers (Figure 3B, i and ii), and fewer than half of URMF postdocs (48%) reported feeling that they belonged socially in their graduate departments.

Postdocs were also asked their research self-efficacy, that is, the extent to which they agreed with the statement “I am confident in my abilities as an independent researcher” (Figure 2C). While more than 60% of postdocs from all social groups reported having confidence in their research abilities, there were again differences by social identity. Men reported higher levels of research self-efficacy than women (WRM: 83%; URMM: 87%), and URMF reported the lowest levels of research self-efficacy (URMF: 63%; WRF: 72%). These group differences in levels of research self-efficacy remained even when accounting for research productivity (i.e., number of first-authored, peer-reviewed publications normalized for amount of time in training; unpublished data).

Different Patterns of Career Interest Development among Male and Female Postdocs Occur in Graduate School

Statistical modeling was used to describe the relationships between these factors and career interests, and to determine whether any of these factors might explain differences in career interest patterns by social identity. Specifically, we used logistic regression analysis, in which the likelihood that a postdoc would express high interest in each career pathway (i.e., answer 4 or 5 on the interest scale) was the outcome/dependent variable. Four models were tested. Each model incorporated social identity, with the career interests of WRM serving as the reference group. Subsequent models incorporated factors such as research productivity and research self-efficacy (model 2), a postdoc’s career interest at PhD completion (model 3), and training experiences (model 4). Results of the models for faculty careers at research universities and nonresearch careers are shown in Tables 1 and 2, respectively (results for faculty at a teaching-intensive university and a research career outside academia are shown in Supplemental Material Tables S9 and S10).

With respect to faculty careers in research-intensive universities, career interests at PhD completion were able to account for postdoc males’ higher levels of interest when compared with females. Among the postdocs in the study, URM postdocs were as likely as WRM to report high interest (odds ratio [OR]: 0.89; 95% confidence interval [CI]: 0.53–1.49; p = 0.66); however, on average, WRF were 39% less likely (OR: 0.61; 95% CI: 0.48–0.76; p < 0.001) and URMF were 65% less likely (OR: 0.35; 95% CI: 0.23–0.53; p < 0.0001) to report high interest (Table 1, model 1). Both research self-efficacy (OR: 2.94; 95% CI: 2.31–3.72; p < 0.001) and research productivity (OR: 3.70; 95% CI: 2.23–6.13; p < 0.001) were positively associated with interest in faculty careers in research-intensive universities (Table 1, model 2). However, after controlling for these factors, WRF remained 30% less likely (OR: 0.70; 95% CI: 0.54–0.89; p < 0.004) and URMF were 54% less likely (OR: 0.46; 95% CI: 0.30–0.70; p < 0.001) than WRM to express high interest in faculty positions at research-intensive universities (Table 1, model 2). In other words, postdoc women with productivity and research self-efficacy similar to their male counterparts were less likely to express interest in a faculty career at a research-intensive institution. When taking into account whether or not a postdoc had high interest in this career path at PhD completion (OR: 41.72; 95% CI: 28.14–61.84; p < 0.001), there were no longer differences in interest by social identity (Table 1, model 3). Thus, our analysis suggests the observed differences in interest by social identity group are explained by women’s lower levels of interest in faculty careers at the end of graduate school. Model 4 showed that interest at PhD completion remained the strongest predictor of high interest in a faculty career currently (OR: 38.43; 95% CI: 23.25–63.49; p < 0.001), even when accounting for other variables such as postdoctoral advisor investment (OR: 1.61; p < 0.05), research self-efficacy (OR: 2.50; p < 0.001), and research productivity (OR: 2.52; p < 0.05).
Thus, in this sample of postdocs, women’s lower levels of interest in faculty careers at research universities at PhD completion explains much of their lower levels of interest as postdocs.

Similarly, the higher interest of URMF in careers outside research relative to other groups (Table 2, model 1) appears to be explained by differences that occur during graduate training. In the regression analyses assessing factors influencing interest in careers outside research, models 3 and 4 show that high interest at PhD completion strongly predicts high interest currently, and, when this is accounted for, there are no longer significant differences among postdocs of different backgrounds. Thus, the higher level of interest of URMF currently is largely explained by their high level of interest at PhD completion.
interest in nonresearch careers at the end of their graduate training. Beyond interest at PhD completion, high levels of interest in careers outside research were inversely associated with postdoctoral advisor career investment (OR: 0.58; 95% CI: 0.43–0.77; \( p < 0.001 \)) and positively associated with longer length of time in postdoctoral training (OR: 1.20; 95% CI: 1.03–1.39; \( p < 0.02 \)). In other words, holding all other factors constant, a fifth-year postdoc was 2.5 times more likely than a first-year postdoc to express high interest in a career outside research. Collectively, these results show that the group differences in the career interests of postdocs can be explained by differences in career interests at PhD completion.

**DISCUSSION**

Postdoctoral scientists are critical to the research enterprise but remain understudied. This survey of American postdocs was done in an effort to illuminate their career development and to inform the ongoing efforts to enhance biomedical training and workforce diversity. The findings of this research have implications for our understanding of postdocs’ knowledge about career options, their diverse patterns of career development, and the factors related to their career choices.

First, these results point to the need for enhanced career development over the full course of scientific training. Our findings suggest that trainees start graduate school with little knowledge about the various careers they could pursue with their degrees. While the increase in the number of participants who feel knowledgeable about their career options throughout their training is significant and notable, clarity about career goals declines during this time. Further, much of what trainees learn before their postdoctoral appointments appears to be outside formal programming in their respective departments. Given that increased access to information does not appear to translate to increased certainty about one’s career path, it is important to consider whether traditional career development opportunities that...
suggests that enhanced career development at the undergraduate level may also be beneficial. Earlier opportunities to reflect critically on career options and interests may be particularly important for trainees from groups underrepresented in the research workforce (women and URM). Recent efforts at broadening participation have focused primarily on enhancing the number of scientists from URM backgrounds who complete PhDs and mentoring to support students from these populations (NIH, 2014a). While the postdocs in this study did not report significant differences in advisor relationships across social groups, there were clearly gendered differences in the career interests. Women were less likely—and URMF particularly less likely—than men to show interest in faculty careers at research universities, and these differences remained when accounting for any differences in research self-efficacy or productivity. Interestingly, among postdoc men, the career interests of the URM largely mirrored those well-represented men—a notable difference from the wider pool of PhD graduates (Gibbs et al., 2014). Racial and gender differences in this career path were no longer significant once level of interest in pursuing a faculty career at a research university at the end of graduate school was controlled. These findings, coupled with the descriptive analyses of participants’ career paths, suggest that provide access to information about options (e.g., seminars and panels) are meeting trainees’ needs. In addition to information, these findings suggest the need for departments and universities to implement more structured opportunities for trainees to reflect critically on what career paths best meet their needs, values, and interests, and to offer more opportunities to begin to explore those career paths (e.g., experiential learning). A number of institutions have begun to do this through mechanisms such as the NIH BEST awards (NIH, 2014b), and these results suggest a need for efforts like this to be expanded.

The findings also suggest a misalignment between when career decisions are made and structured career development is offered. While postdocs’ career interests were largely formed during graduate training, they did not report having been offered structured career development during this time. Instead, their postdoctoral institutions offered structured career development. While career development of postdoctoral scientists will continue to be important, these results reinforce the recommendation of the National Academies postdoc report that “beginning in the first year of graduate school, [host institutions and mentors should] make graduate students aware of the wide variety of career paths available for PhD recipients” (NAS, 2014, p. 4). Moreover, the poor knowledge of career options reported at PhD entry suggests that enhanced career development at the undergraduate level may also be beneficial.

Earlier opportunities to reflect critically on career options and interests may be particularly important for trainees from groups underrepresented in the research workforce (women and URM). Recent efforts at broadening participation have focused primarily on enhancing the number of scientists from URM backgrounds who complete PhDs and mentoring to support students from these populations (NIH, 2014a). While the postdocs in this study did not report significant differences in advisor relationships across social groups, there were clearly gendered differences in the career interests. Women were less likely—and URMF particularly less likely—than men to show interest in faculty careers at research universities, and these differences remained when accounting for any differences in research self-efficacy or productivity. Interestingly, among postdoc men, the career interests of the URM largely mirrored those well-represented men—a notable difference from the wider pool of PhD graduates (Gibbs et al., 2014). Racial and gender differences in this career path were no longer significant once level of interest in pursuing a faculty career at a research university at the end of graduate school was controlled. These findings, coupled with the descriptive analyses of participants’ career paths, suggest that

<table>
<thead>
<tr>
<th>Covariate class</th>
<th>Covariate</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social identity</td>
<td>WRM Reference</td>
<td>1.12 (0.68–1.84)</td>
<td>1.05 (0.59–1.86)</td>
<td>1.08 (0.54–2.18)</td>
<td>1.01 (0.42–2.48)</td>
</tr>
<tr>
<td></td>
<td>URM Reference</td>
<td>1.12 (0.88–1.41)</td>
<td>1.04 (0.81–1.34)</td>
<td>0.97 (0.68–1.37)</td>
<td>1.01 (0.66–1.56)</td>
</tr>
<tr>
<td></td>
<td>URMF Reference</td>
<td>1.81 (1.02–3.21)*</td>
<td>1.51 (0.83–2.76)</td>
<td>1.14 (0.63–2.06)</td>
<td>0.79 (0.42–1.48)</td>
</tr>
<tr>
<td>Personal disposition</td>
<td>High career pathway interest at PhD completion</td>
<td></td>
<td></td>
<td></td>
<td>32.09 (19.11–53.88)**</td>
</tr>
<tr>
<td></td>
<td>Research self-efficacy (i.e., confidence in ability as an independent researcher)</td>
<td>0.66 (0.53–0.80)**</td>
<td></td>
<td></td>
<td>31.62 (19.10–52.36)**</td>
</tr>
<tr>
<td>Objective</td>
<td>First-author publication rate (publications/years in training)</td>
<td></td>
<td></td>
<td></td>
<td>0.51 (0.33–0.78)*</td>
</tr>
<tr>
<td></td>
<td>Total time in postdoctoral training (years)</td>
<td></td>
<td></td>
<td></td>
<td>1.07 (0.55–2.08)</td>
</tr>
<tr>
<td>Postdoctoral training experiences (sense of belonging, advisor support, career development)</td>
<td>Intellectual belonging, research group Reference</td>
<td>0.79 (0.54–1.14)</td>
<td></td>
<td></td>
<td>1.20 (1.03–1.39)*</td>
</tr>
<tr>
<td></td>
<td>Social belonging, research group</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Postdoctoral advisor invested in career advancement</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Postdoctoral advisor equally supportive of students pursuing academic and nonacademic career paths</td>
<td>1.10 (0.85–1.45)</td>
<td>0.58 (0.43–0.77)**</td>
<td>1.03 (0.76–1.39)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Institution offered structured opportunities to explore a variety of career pathways</td>
<td>1.12 (0.81–1.54)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Institution offered structured opportunities to develop nonresearch skills</td>
<td>0.91 (0.65–1.26)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Adjusted odds ratios (and 95% CI) shown. Likelihood of expressing high interest in career path across social identity, adjusted for: model 1: unadjusted; model 2: research self-efficacy and productivity; model 3: high interest in career pathway at PhD completion; and model 4: high interest in career pathway at PhD completion, research self-efficacy, objective measures, and postdoctoral training experiences.

*p < 0.05; **p < 0.001.
much of the career decision making for URMF and WRF scientists takes place during graduate school. We have completed in-depth interviews with nearly 70 survey respondents representing a diversity of social backgrounds (WR and URM men and women) and career pathways (academic and nonacademic). Subsequent analyses will focus on the factors (personal, institutional, and structural) that may be causing these unique patterns of career development such that institutions and federal agencies to can craft policies to make research careers attractive to trainees from all backgrounds.

Enhanced career development earlier in training could help alleviate the so-called glut of postdocs (Davis, 2005; Kaplan, 2012), better equipping trainees to start their careers with greater clarity after PhD completion. The NAS (2014) recommended that postdoctoral positions be “only for those seeking advanced research training” (p. 4) and that they “should not be viewed as the default step after the completion of doctoral training” (p. 5). However, a large percentage of the postdocs in this sample pursued postdoctoral positions, despite having interests in careers outside research. Thus, these data suggest there are many trainees in the current postdoctoral pool who are not preparing exclusively for faculty careers (or even careers in research). These results reinforce the need for biomedical graduate and postdoctoral training to adapt so that trainees can learn about and develop the skills necessary for the full range of careers available, considering both their interests and the realities of the academic job market.

ACKNOWLEDGMENTS

We thank Keith Micoli, Chris Pickett, and Sibby Anderson-Thompson for their thoughtful feedback before submission. This work was supported, in part, by a generous grant from the Burroughs Wellcome Fund (Grant Number 1011798). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. K.D.G. is a member of the board of directors for the National Postdoctoral Association. Findings and views expressed in this work are those of the authors alone and do not necessarily reflect the views of any organization with which they are affiliated.

REFERENCES


Gibbs KD Jr, 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.


Martinez ED, Botos J, Dohoney KM, Geiman TM, Kolla SS, Oliveira A, Qiu Y, Rayasam GV, Stavreva DA, Cohen-Fix O (2007). Falling off the academic bandwagon. Women are more likely to quit at the postdoc to principal investigator transition. EMBO Rep 8, 977–981.


Cues Matter: Learning Assistants Influence Introductory Biology Student Interactions during Clicker-Question Discussions

Jennifer K. Knight,* Sarah B. Wise,† Jeremy Rentsch,*‡ and Erin M. Furtak.§

*Department of Molecular, Cellular, and Developmental Biology, †Department of Ecology and Evolutionary Biology, and §School of Education, University of Colorado Boulder, Boulder, CO 80309

Submitted April 13, 2015; Revised August 15, 2015; Accepted August 17, 2015

The cues undergraduate biology instructors provide to students before discussions of clicker questions have previously been shown to influence student discussion. We further explored how student discussions were influenced by interactions with learning assistants (LAs, or peer coaches). We recorded and transcribed 140 clicker-question discussions in an introductory molecular biology course and coded them for features such as the use of reasoning and types of questions asked. Students who did not interact with LAs had discussions that were similar in most ways to students who did interact with LAs. When students interacted with LAs, the only significant changes in their discussions were the use of more questioning and more time spent in discussion. However, when individual LA–student interactions were examined within discussions, different LA prompts were found to generate specific student responses: question prompts promoted student use of reasoning, while students usually stopped their discussions when LAs explained reasons for answers. These results demonstrate that LA prompts directly influence student interactions during in-class discussions. Because clicker discussions can encourage student articulation of reasoning, instructors and LAs should focus on how to effectively implement questioning techniques rather than providing explanations.

INTRODUCTION

One of the scientific practices most often emphasized in science education reform is argumentation, or the discussion and defense of competing ideas. Prior research has established that engaging students in argumentation can build students’ abilities to understand, practice, and participate in science (Osborne et al., 2004). In addition, argumentation has been shown to encourage scientific thinking, since this process involves students confronting different ideas about content as they describe their reasoning to one another (e.g., Kuhn, 1993; Koslowski, 1996; Zohar and Nehmet, 2002; Asterhan and Schwarz, 2009). Ultimately, students who are taught the principles of argumentation as part of their science courses have the potential to perform better on assessments that require reasoning (Bao et al., 2009; Osborne, 2010). In fact, many recent publications about science education have emphasized the importance of engaging students in such practices to support their learning of science content (e.g., National Research Council [NRC], 2007; American Association for the Advancement of Science [AAAS], 2011). This focus is an extension of decades of efforts at the K–12 level that have sought to bring authentic science practices, critical thinking, and problem solving into classrooms and to engage students in scientific practices intertwined with their learning of scientific ideas (Duschl, 2008; NRC, 2012).

Unfortunately, despite national efforts, incorporating these practices has proven challenging for K–12 teachers.
advanced students engaging in group discussions of clicker questions were sensitive to the language used by the instructor to cue the beginning of their discussion. When students were prompted to use reasoning and were prepared to discuss their reasoning with the rest of the class, they were significantly more likely to engage in the exchange of complete reasoning statements than when they were prompted to simply discuss the right answer. This finding strongly suggests that the students were paying attention to the cue given to them and were changing their interactions with one another dependent on this cue.

A second approach that may ameliorate challenges to encouraging argumentation in science undergraduate courses is the use of peer coaches to facilitate student in-class discussion. Peer coaches can encourage the use of reasoning and argumentation during class and can also guide students in problem solving and other hands-on activities. At the University of Colorado, peer coaches are called learning assistants (LAs), and the learning assistant program has become an international model for training students to serve in this role, primarily in large-enrollment introductory science courses (Otero, 2006). LAs are trained through a pedagogy course in the School of Education that emphasizes techniques to encourage students to construct their own ideas. In physics courses, LAs have been shown to improve student learning gains on concept assessments, and to improve student attitudes toward science (Otero et al., 2010). Students in similar programs, such as peer-led team learning and peer-led guided inquiry, also benefit from interactions with their peer leaders, making significantly higher gains in science critical-thinking skills than their peers (Lewis and Lewis, 2008; Quitadamo et al., 2009). Thus, the peer-coach model may be a way to encourage, model, and support argumentation in class, helping students to develop these skills while they are learning content.

In this paper, we kept constant how instructors cued the students and instead investigated how the presence of LAs might additionally impact student interactions during clicker discussions. We hypothesized that introductory students who were being cued by their instructors to use reasoning in their discussions would regularly use both questioning and reasoning but that the quality and quantity of such interactions would increase in the presence of a peer coach. We first characterize the general features of introductory students’ clicker discussions, only some of which were conducted in the presence of a peer coach. We then characterize the discussions of groups who both did and did not interact with peer coaches and, finally, look more carefully at individual student responses to different kinds of cues from the peer coaches.

METHODS

Course Characteristics
We performed this study in a freshman-level introductory molecular and cell biology course that is required for students planning to major in this discipline and is also taken by students intending other majors (e.g., integrative physiology and neuroscience). Two experienced instructors (not the authors) cotaught two sections of this course, with a total enrollment of ~450 students; because of room and time
constraints, one section was large, and one was much smaller (94 students). This study was carried out in the smaller section of 94 students. The instructors spent class time (50 min, three times per week) engaged in lecture interspersed with three to five clicker questions per class period, using the iClicker response system, with time given for discussion and feedback on each question. The following clicker-question cycle with peer discussion (Mazur, 1997) was used most frequently: the instructor displayed the clicker question and asked students to make an initial independent vote; the instructor then cued students to enter into discussion in their small groups, reminding them to use reasons to back up their preferred answers; students revoted, and at the conclusion of the voting, the instructor usually called for volunteers to explain their reasoning before showing the histogram of votes. These cues closely resemble the “reasoning-centered” approach used in our previous study of advanced students (Knight et al., 2013). A slight modification to this procedure was used in about one-third of the discussions: the instructor posed the question and asked students to think about it on their own but did not record an official individual vote before allowing students to move on to discussion. For these questions, students only submitted a postdiscussion vote.

**Study Participants**

All students in the course were asked to self-select into groups of three to four students for in-class discussion and to sit with their groups in each class period. Out of the 94 students in the lecture section studied, 23 volunteered to participate in this study (six groups). This method of participant selection was chosen for several reasons: volunteer students agreed to remain in these groups and sit in the same place in each class period, and they were willing to volunteer for the duration of the study, making it feasible to find the volunteer students in class and obtain reliable recordings. In addition, multiple recordings of several groups of students, rather than single recordings of a larger number of groups of students, allowed us to control for clicker-question variation and group-specific discussion variation in our analyses, since we collected multiple discussions for each group and multiple discussions of each clicker question from different groups.

**Role and Participation of LAs**

The LAs’ primary role in this course was to lead problem-solving sessions outside class, with a secondary role of attending class and interacting with students during clicker discussions. LAs were either currently enrolled in or had recently completed an LA pedagogy course; they also met with the course instructors once per week to review content. For the purposes of this study, the LAs were asked to apply the training from their pedagogy course to facilitate in-class clicker discussions but were not given any additional explicit instructions. LAs were assigned to sit with a specific study volunteer group during one class period and with a non-volunteer group during the next class period; in this way, study volunteers interacted with an LA during about half of the class periods. Because only three LAs could regularly attend this section, only three of the six volunteer groups had recurring audio recordings with and without an LA. The other three groups did not encounter an LA regularly during the period of this study, although it is possible that they may have interacted with LAs outside the class periods we recorded or in ways we could not capture with our audio devices.

**Data Collection**

Each of the volunteer students wore a wireless microphone (lavalier style) during eight class periods (weeks 8–10 of a 15-wk semester). Demographic information provided by the registrar’s office established that the volunteer students had significantly higher incoming GPAs than the section average but were otherwise representative of their section with respect to gender distribution and year in school (Table 1).

We used a Nady receiver and a digital audio recorder (Zoom Corporation) to combine wirelessly transmitted audio from each volunteer group of students during their discussions of clicker questions. The audio recordings were transcribed into an Excel spreadsheet, paired with the clicker question the students had discussed, and given a unique transcript number. Each speaker was given a number within a transcript to facilitate tracking student interactions and to tally the number of speakers per discussion. However, individual speaker identifications within a group could not be reliably preserved from discussion to discussion, precluding us from identifying and following an individual’s specific contribution over time.

The time given by the instructor for each discussion and time spent in on-task discussion (Discussion Length) was noted for each recording, and Percent Productivity was calculated from these two numbers (discussion length/time given). Small deviations from the task, such as one off-topic statement, were ignored, but if students were off-topic for 30 s or more, that time was subtracted. Transcripts were then coded for features of discussion (described in Data Analysis), and these features were summed across discussions.

The data set for this study includes 140 discussions by six groups of students on 28 clicker questions. Owing to student absences and occasional problems with recording equipment, no single group of students was recorded discussing all 28 questions. The number of questions discussed by each group ranged from 17 to 27, as shown in Table 2. Because students voted twice on only a subset of clicker questions, this subset of 83 discussions across all groups was used to describe the impact of discussion on performance (initial vote to revote; Table 2). To explore the effect of LA presence on student discussion, we used 65 transcripts of three groups of students (12 students total) who were audio-recorded in both the presence and absence of LAs. The remaining

### Table 1. Demographics of students’ class rank

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>% Female</th>
<th>Class rank</th>
<th>GPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonvolunteers</td>
<td>71</td>
<td>46</td>
<td>1.7 (0.9)</td>
<td>2.8 (0.7)</td>
</tr>
<tr>
<td>Volunteers</td>
<td>23</td>
<td>52</td>
<td>1.8 (0.9)</td>
<td>3.2 (0.8)</td>
</tr>
</tbody>
</table>

*a Freshman = 1; sophomore = 2, etc. The volunteers are no different from the rest of the students in gender or class rank (p < 0.05, Mann-Whitney U-test).

*b Volunteer GPA is significantly higher than nonvolunteers (p < 0.05, t-test).
Discussions were coded in two ways: global, and line by line. In global coding, the transcript as a whole was characterized by the presence of each discussion characteristic. In this phase of coding, the unit of analysis was the group’s whole discussion of one clicker question rather than individual student statements. LA statements were not counted as contributing to these global codes (Table 3). However, during the second phase of line-by-line coding, we characterized individual statements made by LAs and the responses made by students to each LA statement. This coding is described in more detail below in Student Responses to Different Types of LA Statements.

All discussion codes were developed using an iterative process, building upon our experience coding advanced student discussions (Knight et al., 2013) and using a system based on Toulmin’s characteristics of argumentation (Toulmin, 1958). We read through many student exchanges, discussed the interactions students engaged in, and settled on categories that were descriptive and potentially interesting. We chose to follow two global codes described previously in Knight et al. (2013): Exchange of Quality Reasoning and Conflicting Reasoning (which we modified into Reasoning about Multiple Answers) (Table 3). Exchange of Quality Reasoning characterizes reasoning and the use of “warrants” (Toulmin, 1958) as well as whether an exchange of such reasoning is occurring. Warrants are complete reasoning statements, in which a student provides a reason for his or her answer and connects this reason logically to data or factual information. Students can also articulate a less complete form of reasoning in which they may suggest

### Table 2. Distribution of recorded discussions among different volunteer groups

<table>
<thead>
<tr>
<th>Groups who interacted with an LA</th>
<th>Number of discussions recorded</th>
<th>Number of discussions with initial and revote</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>24</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>24</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>65</td>
<td>36</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Groups who did not interact with an LA</th>
<th>Number of discussions recorded</th>
<th>Number of discussions with initial and revote</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>27</td>
<td>17</td>
</tr>
<tr>
<td>5</td>
<td>21</td>
<td>13</td>
</tr>
<tr>
<td>6</td>
<td>27</td>
<td>17</td>
</tr>
<tr>
<td>Total</td>
<td>75</td>
<td>47</td>
</tr>
</tbody>
</table>

Total for both groups 140 83

75 transcripts from three groups who never interacted with LAs (11 students total) were used as a comparison group.

Finally, each clicker question was rated by two people not associated with the study as either higher-order or lower-order cognitive level, using Bloom’s taxonomy (Anderson and Krathwohl, 2001; Crowe et al., 2008). Raters agreed on 73% of Bloom’s ratings and adjudicated all differences. Sixteen questions were rated as higher order and 12 as lower order.

### Table 3. Description of global codes

<table>
<thead>
<tr>
<th>Global code</th>
<th>Definition/Characteristics</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exchange of Quality Reasoning (0–3)</td>
<td>No reason provided</td>
<td>“What did you vote?” “A.” “I think it’s because of transcription being different.” “Yeah.”</td>
</tr>
<tr>
<td>0</td>
<td>One person provides reason(s)</td>
<td>“I think it’s because transcription is different in eukaryotes and prokaryotes.” “Yeah, and because of the sigma factor …”</td>
</tr>
<tr>
<td>2</td>
<td>Two or more people provide simple reason(s)</td>
<td>“I think it’s because … there’s no nucleus in bacteria, so that would be a difference between eukaryotes and bacteria.”</td>
</tr>
<tr>
<td>3</td>
<td>Two or more people provide reasons supported by evidence and a logical connection (warrants)</td>
<td>“Yes, there’s no need to transport the transcript out of the cytoplasm since the enzyme for making the mRNA transcript is right there.”</td>
</tr>
<tr>
<td>Reasoning about Multiple Answers</td>
<td>More than one answer is considered, using reasoning</td>
<td>“It doesn’t have anything to do with the membrane [answer C] because …” “But I think the concentration [answer A] does matter because …”</td>
</tr>
<tr>
<td>Hedging a Reason</td>
<td>Signaling uncertainty in one’s own reasoning</td>
<td>“I don’t know, really, but it could be because …” “I think it works this way but I’m totally guessing”</td>
</tr>
<tr>
<td>Analogy or Example</td>
<td>Using an analogy or an example to help explain a reason</td>
<td>“It’s like spraying perfume in a room.”</td>
</tr>
<tr>
<td>Student–Student Questioning</td>
<td>Asking for votes or basic information, like definitions</td>
<td>“What did you vote?”</td>
</tr>
<tr>
<td>Requesting Information</td>
<td></td>
<td>“What does that mean?”</td>
</tr>
<tr>
<td>Requesting Reasoning</td>
<td>Asking to share an explanation</td>
<td>“Why did you say that?” “Why were you thinking that?”</td>
</tr>
<tr>
<td>Requesting Feedback</td>
<td>Asking for confirmation of own reasoning</td>
<td>“It takes energy to break bonds, right?”</td>
</tr>
</tbody>
</table>

*Each discussion was given a 0/1 (absence/presence) for each code, except Exchange of Quality Reasoning, as shown.*
an idea or use a “because”-type statement, typically providing partial evidence supporting an idea but lacking a logical connection to their claim. We were also particularly interested in the kinds of questions that students ask one another as they discuss their ideas; the question codes we ultimately developed were based on “talk moves” used by effective K–12 teachers to prompt student discussion (Michaels and O’Connor, 2012). We iteratively refined the definition of each code as we practiced coding transcripts until the definition was clear and could be reliably recognized by multiple coders.

To establish interrater reliability, three raters each coded the same set of transcripts, discussing results and adjudicating differences over several training sessions. Interrater reliability was then established using intraclass correlation between these three raters on 24% of the total number of transcripts, achieving a Cronbach’s alpha greater than 0.75 for each code reported. The line-by-line codes used to describe specific LA statements and the responses of students to those statements were developed using a similar process. These codes describe similar interactions to the global codes (e.g., reasoning, different kinds of questioning) but are used to describe each speaker’s contribution to the discussion (described in more detail in Student Responses to Different Types of LA Statements). Two raters carried out the line-by-line coding of the entire set of transcripts independently with 86% agreement and came to consensus on any differences.

Regression Analyses
SPSS version 22 was used to run multiple regression analyses (linear, binary logistic, and ordinal logistic) to determine how each possible variable impacted the global outcome codes of the transcripts. Linear regressions were conducted for the continuous outcomes of Discussion Length and Percent Productivity. Binary logistic regressions were conducted for all presence/absence outcomes, and ordinal logistic regressions were conducted for Exchange of Quality Reasoning. The regression models generally used the following covariates: group ID, Bloom’s level of question, number of speakers, and study day. Study day was assigned based on the actual day of class (i.e., class 8, 10, 17, etc.) to account for any differences over the recording time frame. An additional possible covariate was whether questions were discussed with only one vote following discussion or with an initial vote followed by discussion and revote. As this factor had no statistically significant effect on any outcomes and did not affect the significance of any models, it was not included in the final regression analyses. For linear regression models, the assumptions of linearity, independence of errors, homoscedasticity, unusual points, and normality of residuals were all met and produced models with p values < 0.01. For logistic regressions, there were no significant interactions between the continuous outcome variables, thus meeting the assumption of linearity for each analysis. For ordinal regressions, multicollinearity, proportional odds, goodness of fit, and model fitting were all within acceptable parameters (Field, 2009).

Human Subjects Approval
This work was reviewed by the University of Colorado Institutional Review Board, and the use of human subjects was approved (expedited, protocol #11-0630).

RESULTS
We first describe general student discussion characteristics and performance, then the comparisons of groups who had access to an LA during their discussions, and, finally, a detailed analysis of student responses to specific types of LA prompts.

Discussion Characteristics
Students generally began discussing a clicker question within a few seconds of being prompted by the instructor and stayed on task most of the time. Sometimes, students exchanged off-topic talk before they began discussion on the clicker question, paused in their discussion, or, in a few instances, became derailed in the middle of their discussion by a humorous remark or someone telling a story. In most cases, students discussed a topic until they reached a decision, at which point they might engage in social talk. The total time given by an instructor to any discussion depends on factors such as how quickly student votes were recorded, or whether the instructor had a conversation with a group of students during the voting period, and thus let the discussion continue for longer than normal. Overall, students spent an average of 1.25 min (± 0.68) in discussion, amounting to 64% of the time the instructor allotted (2 min ± 0.86). Notably, students engaged in discussions relevant to the clicker question 93% of the time.

To understand more about the nature of student discussion, we characterized each transcript, using seven global codes that reveal how students use reasoning and questioning: Exchange of Quality Reasoning, Reasoning about Multiple Answers, Hedging a Reason, using Analogy or Example, Requesting Information, Requesting Reasoning, and Requesting Feedback (each described in detail in Table 3). We calculated the frequency with which each of these global codes appeared in the data set of 140 discussions (Table 4). Overall, students used some kind of reasoning (levels 1–3) in 91% of their discussions but only exchanged warrants (level 3) 18% of the time. Many discussions (42%) contained a level 2 Exchange of Quality Reasoning, in which students exchanged reasons for their ideas but not necessarily warrants. A second measure of reasoning was Reasoning about Multiple Answers, in which reasons for more than one answer were considered by the group. This characteristic correlated with the level of Exchange of Quality Reasoning achieved in the discussion: in level 1 discussions, only 35% used Reasoning about Multiple Answers; thus, only one person used reasoning but offered or explained reasoning about multiple possible answers to the rest of the group. In level 2 discussions, 75% included reasoning about multiple answers, and in level 3 discussions, 100% included this type of exchange.

Two additional discussion characteristics were relatively infrequent: Hedging a Reason and using Analogy or Example. Hedging a Reason, in which students softened their reasoning with qualifiers such as “that’s what I think, but I could be wrong,” occurred in only 14% of discussions. Using an Analogy or Example, for example, likening the diffusion of ions to the diffusion of perfume, was similarly infrequent. Although these characteristics are potentially interesting, we chose not to follow them further with regression analyses.
The last set of characteristics described how students used questioning in their discussions. Questioning was used in almost every discussion, and the three types of questions we coded were frequently used in combination: 94% of discussions included more than one type of question, with Requesting Information the most common, followed by Requesting Feedback, and then Requesting Reasoning (Table 4). A typical sequence of questions in a discussion often followed this pattern: one or more information requests, each followed by statements of the answers chosen (e.g., “What did you vote?” “C.” “What about you?” “A.”), then a direct request for reasoning (“Why did you think it was C?”), and a statement of a reasoning that sometimes included a request for feedback (“It is A because of [reasoning statement], right?”). Such an exchange often would end with students agreeing that the reason made sense, but sometimes another round of questioning would ensue if an alternative idea was proposed.

Table 4. Frequency of global codes in student discussions

<table>
<thead>
<tr>
<th>Whole discussion codes</th>
<th>Total frequencies (all 140 discussion)</th>
<th>Frequency of each behavior (%) among groups who interacted with LAs (n = 65 discussions)</th>
<th>Frequency of each behavior (%) among groups who did not interact with LAs (n = 75 discussions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exchange of Quality Reasoning</td>
<td>0</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>31</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>42</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>18</td>
<td>14</td>
</tr>
<tr>
<td>Reasoning about Multiple Answers</td>
<td>61</td>
<td>57</td>
<td>65</td>
</tr>
<tr>
<td>Hedging a Reason</td>
<td>14</td>
<td>12</td>
<td>16</td>
</tr>
<tr>
<td>Analogy or Example</td>
<td>11</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>Questioning</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Requesting Information</td>
<td>66</td>
<td>74</td>
<td>60</td>
</tr>
<tr>
<td>Requesting Reasoning</td>
<td>29</td>
<td>50</td>
<td>22</td>
</tr>
<tr>
<td>Requesting Feedback</td>
<td>41</td>
<td>51</td>
<td>32</td>
</tr>
</tbody>
</table>

Clicker-Question Performance

During the study period, students answered 28 clicker questions. In 20 of these, students submitted both an initial vote and a revote following discussion. Altogether, students averaged 50% correct on their initial votes and 71% correct on their revotes. In comparing volunteer students with the rest of the students in the course, both volunteers and nonvolunteers improved in their average percent correct from initial to revote, and there is no significant difference in the two groups for either initial or revote (p = 0.45; Figure 1). In addition, we note that the average percent correct for questions on which there was no initial vote (77%) was similar to that of questions for which there was both an initial and revote.

To determine the potential factors affecting students’ percent correct on revotes, we looked at the subset of discussions for which we had both initial and revotes recorded, and more than one person per group responding. For these 81 discussions of 20 questions, we could calculate the initial and revote percent by group and determine whether the students were in agreement on their initial vote (irrespective of correctness). Before discussion, students were in unanimous agreement on their initial vote only 25% of the time. Using a linear regression with initial agreement (0/1), initial percent correct by group, group ID, Bloom’s level of the question, number of speakers, Discussion Length, and Ever with LA as covariates, we found only the initial percent correct significantly predicted the percent correct on the revote (beta = 0.43; F(7, 73) = 3.52, p < 0.005). In this model, ~20% of the variance was accounted for by the independent variables (adjusted $R^2 = 0.19$). We also explored whether any of the coded discussion features (such as Exchange of Quality Reasoning) could predict higher percent correct on revotes; these linear regression models had adjusted $R^2 = 0.18–0.23$, but no discussion features predicted the percent correct (except the covariate initial percent correct, described earlier).

Differences between Groups Who Regularly Interacted with LAs and Those Who Did Not

Because we recorded three groups (11 students) who never interacted with LAs during our recording sessions, and an
Table 5. Regression table for all 140 discussions

<table>
<thead>
<tr>
<th>Regression factors</th>
<th>Exchange of Quality Reasoning</th>
<th>Reasoning about Multiple Answers</th>
<th>Requesting Information</th>
<th>Requesting Reasoning</th>
<th>Requesting Feedback</th>
<th>Discussion Length</th>
<th>Percent Productivity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Odds ratio (p value)</td>
<td>Odds ratio (p value)</td>
<td>Odds ratio (p value)</td>
<td>Odds ratio (p value)</td>
<td>Odds ratio (p value)</td>
<td>Beta (p value)</td>
<td>Beta (p value)</td>
</tr>
<tr>
<td>Ever with LA</td>
<td>0.52 (0.27)</td>
<td>0.67 (0.29)</td>
<td>1.78 (0.15)</td>
<td>2.53 (0.03)*</td>
<td>1.97 (0.07)</td>
<td>0.11 (0.17)</td>
<td>0.20 (0.01)*</td>
</tr>
<tr>
<td>Group ID (high)</td>
<td>(0.39)</td>
<td>1.03 (0.68)</td>
<td>0.94 (0.43)</td>
<td>1.13 (0.16)</td>
<td>1.15 (0.07)</td>
<td>-0.08 (0.34)</td>
<td>-0.05 (0.54)</td>
</tr>
<tr>
<td>Bloom’s level</td>
<td>1.08 (0.82)</td>
<td>0.49 (0.67)</td>
<td>1.07 (0.87)</td>
<td>0.94 (0.88)</td>
<td>0.64 (0.24)</td>
<td>0.03 (0.70)</td>
<td>-0.14 (0.06)</td>
</tr>
<tr>
<td>Number of speakers</td>
<td>1.67 (0.02)*</td>
<td>1.77 (0.02)*</td>
<td>2.09 (0.00)*</td>
<td>2.59 (0.00)*</td>
<td>1.67 (0.03)*</td>
<td>0.39 (0.00)*</td>
<td>0.29 (0.00)*</td>
</tr>
<tr>
<td>Study day</td>
<td>0.94 (0.00)*</td>
<td>0.93 (0.00)*</td>
<td>1.05 (0.02)*</td>
<td>0.95 (0.04)*</td>
<td>0.99 (0.58)</td>
<td>-0.23 (0.01)</td>
<td>0.31 (0.00)*</td>
</tr>
</tbody>
</table>

*Odds ratios and p values are shown for each factor’s impact on the coded discussion features shown, using ordinal, logistic, or linear regressions. Asterisks (*) and bold type indicate significant p values.

For logistic regressions, odds ratios < 1 indicate an inverse relationship. The linear regression models were both significant (F(5, 134) = 10.4, p < 0.001), with adjusted R² values of 0.14 for Discussion Length and 0.25 for Percent Productivity.

For independent variables with more than two groups in an ordinal regression, an odds ratio cannot be calculated for the variable, only for each individual group. Instead, the impact of the variable can be represented by the Wald statistic: in this case, Wald χ²(5) = 4.08. The p value for this analysis suggests that group ID does not have a statistically significant effect on the prediction of use of higher level of reasoning.

### Impact of LAs on Discussion

We focused the rest of our analyses on the subset of 65 discussions from the three groups who interacted with an LA. As an LA was available to any group during only about half of the recorded class sessions, this data set allowed us to estimate the impact of LAs on student discussion while controlling for group.

LA presence significantly influenced four discussion outcomes: Requesting Information, Requesting Feedback, Discussion Length, and Percent Productivity. In the presence of LAs, student discussions were 5.6 times more likely to use Requesting Information and 3.9 times more likely to use Requesting Feedback than in the absence of LAs, all else being equal (Table 6). These discussions were also significantly longer (1.58 min ± 0.6) than those without LA interaction (0.96 min ± 0.52 and more productive (78%) than those in the absence of LAs (63%), all else being equal (Table 6). In addition, more discussions involved an exchange of warrants (level 3: 20% vs. 10%), but this difference was not significant.

Some additional covariates were also significant predictors of certain discussion features. Students were significantly more likely to include Reasons for Multiple Answers when the clicker question was lower-order Bloom’s rather than higher-order Bloom’s. The number of speakers was a significant predictor of higher use of Requesting Information and Requesting Reasoning, and Group ID significantly predicted higher use of Exchange of Quality Reasoning and Requesting Feedback (all shown in Table 6). Finally, we also found that study day significantly affected several outcomes, sometimes positively and sometimes negatively (Table 6).

### Student Responses to Different Types of LA Statements

The above analyses indicate that LAs had few significant effects on global features of student discussions. However, if LAs were using a variety of prompting statements when interacting with students, these different prompts could result in different student discussion characteristics. More detailed analysis of LA–student interactions showed this to be the case. Two different discussions of the same clicker question are shown in Figure 2 to illustrate the different responses of students to different LA prompts. In example A, the LA’s initial interaction with students is to provide a reasoning statement, which does not prompt student interaction, while in example B, the LA uses several question prompts to draw out student ideas, resulting in students exchanging ideas.

To better understand how LA prompting statements impacted student responses, we characterized both using the line-by-line codes shown in Table 7. LA statements were coded into five categories: Prompting Questions (asking students which answer they selected or asking them to consider some additional piece of information), Requesting Reasoning (asking questions to elicit reasoning), Background Statement (a statement of factual information but not a reason), and Providing Reasoning (explaining a reason for an answer).
Table 6. Regression table for conversations with an LA (n = 33) and without an LA (n = 32) for the three groups who interacted with LAsa

<table>
<thead>
<tr>
<th>Regression factors</th>
<th>Exchange of Quality</th>
<th>Reasoning about Multiple Answersb</th>
<th>Requesting Informationb</th>
<th>Requesting Reasoninga</th>
<th>Requesting Feedbackb</th>
<th>Discussion Lengthc</th>
<th>Percent Productivityc</th>
</tr>
</thead>
<tbody>
<tr>
<td>LA present</td>
<td>Odds ratio (p value)</td>
<td>Odds ratio (p value)</td>
<td>Odds ratio (p value)</td>
<td>Odds ratio (p value)</td>
<td>Odds ratio (p value)</td>
<td>0.46 (0.00)*</td>
<td>0.24 (0.05)*</td>
</tr>
<tr>
<td>Group ID (high)</td>
<td>(0.05)d</td>
<td>1.18 (0.11)</td>
<td>0.86 (0.19)</td>
<td>1.13 (0.22)</td>
<td>1.24 (0.03)*</td>
<td>−0.03 (0.83)</td>
<td>−0.07 (0.55)</td>
</tr>
<tr>
<td>Bloom’s level (high)</td>
<td>0.72 (0.49)</td>
<td>0.18 (0.01)*</td>
<td>1.57 (0.51)</td>
<td>0.82 (0.74)</td>
<td>0.74 (0.60)</td>
<td>0.05 (0.65)</td>
<td>−0.01 (0.35)</td>
</tr>
<tr>
<td>Number of speakers</td>
<td>1.35 (0.36)</td>
<td>0.96 (0.93)</td>
<td>3.94 (0.01)*</td>
<td>2.72 (0.02)*</td>
<td>0.72 (0.39)</td>
<td>0.14 (0.26)</td>
<td>0.16 (0.22)</td>
</tr>
<tr>
<td>Study day</td>
<td>0.95 (0.08)</td>
<td>0.93 (0.04)*</td>
<td>1.07 (0.08)</td>
<td>0.93 (0.05)*</td>
<td>1.02 (0.74)</td>
<td>−0.21 (0.06)</td>
<td>0.39 (0.001)*</td>
</tr>
</tbody>
</table>

a ODDS ratios and p values are shown for each of the factor’s effects on the characteristics of student discussion using ordinal, binary logistic, or linear regressions. Asterisks (*) and bold type indicate significant p values.

b For logistic regressions, odds ratios < 1 indicate an inverse relationship. Thus, in each of these cases, the odds ratios can be inverted to better describe the outcome: for example, lower-level Bloom’s questions are 5.6 times more likely than higher-level questions to generate reasoning about multiple answers, all else being equal; and the absence of an LA is 5.3 times more likely to generate requests for information, all else being equal.

c For linear regressions, a negative beta value indicates an inverse relationship. The linear regression models were both significant (F(5, 59) = 4.6, p = 0.001), with adjusted R² values of 0.25 for Discussion Length and 0.22 for Percent Productivity.

d For independent variables with more than two groups in an ordinal regression, an odds ratio cannot be calculated for the variable, only for each individual group. Instead, the impact of the variable can be represented by the Wald statistic: in this case, Wald χ²(2) = 5.98. The p value for this analysis suggests that group ID has a statistically significant effect on the prediction of use of higher level of reasoning.

Clicker question: Given what you know about the chemical properties of the lipid bilayer, which of the following proteins is UNLIKELY to be found associated with the membrane?

a. A protein with one or more stretches of ~20-30 consecutive hydrophobic amino acids
b. A globular protein with nonpolar amino acid chains folded into the interior of the protein and hydrophilic side chains exposed to water
c. A protein with 20-30 amino acid stretch of alternating polar and nonpolar amino acids
d. A protein consisting of ONLY hydrophilic amino acids (no hydrophobic amino acids)
e. None of the above.

LA provides reasoning

1. What did you think it was?
2. I’m saying C, a protein consisting of only hydrophobic...

LA: So think about... associating with the membrane can mean within the membrane and it can also be outside the membrane, right? So if it’s associating with the membrane it can be a transmembrane protein and it can also be a peripheral membrane protein, ‘cause it’s still interacting with the membrane.

2. Oh, so I am saying E, it’s none of them.

LA: Yeah, I like E.

LA asks prompting questions

1. So it’s not A, because that can’t exist. B seems like it can exist...
I... don’t know, well I mean it’s not the answer. Let’s see, hydrophilic interactions... you can’t have C it would just be outside of the membrane, I think.

2. I don’t think the answer is D.

LA: You want to think about what proteins do when they associate with the membrane, so... what does D describe?

2. Ummm... A trans protein, well when it goes in and out of the cell... I don’t remember what that’s called.

LA: Yeah and what’s...

1. Well, D is not long enough to go in and out several times, cause it’s only one stretch of 20-30 and that’s like one membrane thick, right? So, E.

or Acknowledgment (a simple statement, such as “yes”). Student responses were also coded into five similar but not identical categories. Because students did not directly ask LAs to explain their reasoning, we combined the two questioning categories into a single category (Asking Questions), and an End of Discussion category was added to track circumstances in which an LA statement was followed by no substantive student contributions. Otherwise, the categories of Using/Providing Reasoning, Background Statements, and Acknowledgment were the same as for the LA statements.

LAs engaged with students on average three times per discussion, for a total of 110 LA statements across the 33 discussions. The most common type of LA statement was a Prompting Question, asking students to share what they voted or to use information they had not yet considered (Table 8). This prompt was typically used by LAs to initiate discussion with the students: in one-third of the discussions, the LAs used a sequence of at least two questions before providing reasoning or background. LAs also frequently used background statements to provide additional information in response to student questions or to prompt additional thinking. In another third of the discussions, the LAs asked a combination of one or more questions and other statements, and then offered their own reasoning. Although LAs were...
Table 7. Line-by-line codes used to describe individual LA statements and the responses of students

<table>
<thead>
<tr>
<th>LA statement</th>
<th>Definition /characteristics</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prompting Question</td>
<td>Request for a student’s answer; request for information</td>
<td>“What did you answer?”</td>
</tr>
<tr>
<td>Requesting Reasoning</td>
<td>Request for sharing an explanation or otherwise providing support for a claim</td>
<td>“What do you think the differences are between those two?”</td>
</tr>
<tr>
<td>Providing Reasoning</td>
<td>Provides explanatory statement of backing, evidence, or justification</td>
<td>“Why did you pick C and not D?”</td>
</tr>
<tr>
<td>Background Statement Acknowledgment</td>
<td>Shares basic information or states support of an answer</td>
<td>“Your genetic code is consistent throughout all organisms … all that really matters is the gene itself.”</td>
</tr>
<tr>
<td>Acknowledgment/Claim</td>
<td>Acknowledgment of a statement</td>
<td>“The genetic code just says UAC codes for tyrosine …”</td>
</tr>
<tr>
<td>Asking Question</td>
<td>Request for basic information, an explanation, or confirmation of reasoning</td>
<td>“Yes.”</td>
</tr>
<tr>
<td>Using Reasoning</td>
<td>Provides explanatory statement of backing, evidence, or justification</td>
<td>“I think the sigma factor is for eukaryotes.”</td>
</tr>
<tr>
<td>Background Statement</td>
<td>Student sharing basic information or states support of an answer</td>
<td>“Yes.”</td>
</tr>
<tr>
<td>Acknowledgment/Claim</td>
<td>Acknowledgment of a statement or statement of a clicker vote choice</td>
<td>“I picked C.”</td>
</tr>
<tr>
<td>End of Discussion</td>
<td>When statement of affirmation ends the discussion</td>
<td>“Okay, that makes sense.”</td>
</tr>
</tbody>
</table>

Student response

<table>
<thead>
<tr>
<th>LA statement</th>
<th>Definition /characteristics</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asking Question</td>
<td>Request for basic information, an explanation, or confirmation of reasoning</td>
<td>“What does that word mean?”</td>
</tr>
<tr>
<td>Using Reasoning</td>
<td>Provides explanatory statement of backing, evidence, or justification</td>
<td>“Why do you think that?”</td>
</tr>
<tr>
<td>Background Statement</td>
<td>Student sharing basic information or states support of an answer</td>
<td>“It takes energy to break bonds, right?”</td>
</tr>
<tr>
<td>Acknowledgment/Claim</td>
<td>Acknowledgment of a statement or statement of a clicker vote choice</td>
<td>See Exchange of Quality Reasoning for examples</td>
</tr>
<tr>
<td>End of Discussion</td>
<td>When statement of affirmation ends the discussion</td>
<td>“Yes.”</td>
</tr>
</tbody>
</table>

encouraged in their general pedagogical training to question and prompt rather than explain the reason for an answer, LAs contributed reasoning statements in more than half of the discussions. However, in only one discussion did an LA begin by providing reasoning. Finally, the least common LA interaction beside simple acknowledgment was directly requesting students’ reasoning.

Each LA statement produced an average of 1.4 student responses, with a total of 156 student responses to the 110 LA statements. To characterize student responses to LA statements, we tracked which student responses followed each type of LA statement until an LA spoke again. For example, in response to Prompting Questions from LAs, student responses included 20 questions, 10 reasoning statements, six background statements, 17 acknowledgments, and one end of discussion. Because acknowledgment statements, although common, do not contribute substantively to the discussion, we did not follow this category further.

The frequency with which each type of LA statement was followed by each student response is shown in Figure 3.

Table 8. Frequency of use of the five types of LA statements*

<table>
<thead>
<tr>
<th>LA statement category</th>
<th>Percent of 33 discussions in which LA statement was used</th>
<th>Percent of 110 total LA statements</th>
<th>Number of student responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prompting Question</td>
<td>64</td>
<td>30</td>
<td>53</td>
</tr>
<tr>
<td>Request for Reasoning</td>
<td>39</td>
<td>17</td>
<td>24</td>
</tr>
<tr>
<td>Using Reasoning</td>
<td>52</td>
<td>25</td>
<td>35</td>
</tr>
<tr>
<td>Background Statement</td>
<td>59</td>
<td>21</td>
<td>36</td>
</tr>
<tr>
<td>Acknowledgment</td>
<td>21</td>
<td>7</td>
<td>9</td>
</tr>
</tbody>
</table>

*More than one type of LA statement was used in 29 of 33 discussions. The total number of student responses to each type of LA statement is also shown.

DISCUSSION

Previous research, primarily in K–12 (Jimenez-Aleixandre et al., 2000; Sandoval, 2003) but increasingly in undergraduate education (e.g., James and Willoughby, 2011; Aydeniz et al., 2012; Kulatunga et al., 2013), has shown the importance of discussion and argumentation among students as a factor in helping them learn both scientific content and the practice of science. Even when discussions do not directly improve short-term performance on test items, exchanging ideas contributes to understanding the discourse practices of science and promotes learning by giving students feedback on their internal construction of ideas and a chance to hear how others are constructing their ideas (Ford and Forman, 2006; Next Generation Science Standards, 2013). In this paper, we have characterized how students discuss clicker questions in a large introductory biology course, with a particular focus on reasoning and questioning, and explored whether interactions with LAs affected student discussion patterns.
We found that LAs can positively influence the articulation of reasoning in student discussions, especially if they use prompting questions and requests for reasoning.

**Student Articulation of Reasoning**

In this study, students were cued by their instructors to discuss the reasons for their answers with one another, and they did so most of the time. However, students relatively infrequently exchanged reasoning with fully articulated warrants—Exchange of Quality Reasoning level 3—claims logically connected to evidence (Table 4). Because these students did not receive explicit instruction on how to use reasoning or on how to construct an argument, this base level of exchange likely is representative of “untrained” student tendencies. Several other studies have shown (primarily with middle and high school students) that discerning what information is relevant in answering a question, using evidence to back up ideas, and justifying one’s reasons are all challenging endeavors (e.g., Erduran et al., 2004; McNeill and Krajcik, 2007). Students do not typically employ these behaviors without training, although they may learn to do so when they frequently work in groups (Lubben et al., 2009).

One recent study of undergraduates showed that students are able to engage in higher levels of reasoning when given complex problem-solving opportunities in a format that explicitly encourages group problem solving and argumentation. Using a modified version of Toulmin’s characterization of argumentation to describe instances of students coconstructing reasoning, Kulatunga et al. (2013) studied a small group of General Chemistry 1 students in peer-led guided inquiry sessions and found that they used coconstructed warrants 50% of the time, considerably higher than the 18% in our sample. This difference may be generated by additional instruction in how to engage in argumentation, differences in the kind of tasks the students were engaged in, or the time given for discussion. These findings support our suggestion that instructional design and practice can substantially influence student use of argumentation.

In comparing different levels of undergraduate biology students, introductory and advanced students clearly differ. Advanced students have engaged in more years of course work and have likely practiced the use of reasoning, even if they have not explicitly been taught to use argumentation. In comparing the introductory students in this study with the advanced students from our previous study, all of whom received a similar instructor cue to consider the reasons for their answers, advanced students clearly use more level 3 exchanges of quality reasoning (more than 50% of discussions; Knight et al., 2013), versus 18% for the introductory students in the current study. While advanced students used predominantly level 3 (exchanges of warrants), introductory students used mostly level 2, in which they exchanged reasons that lacked either the evidence or logical connection to qualify as warrants. Introductory students also used more level 1 reasoning (one student explaining his or her reasoning to others) than did advanced students. This may reflect, in part, the more heterogeneous preparation of introductory students. Some introductory students may feel more knowledgeable or actually be more knowledgeable than others; these students may be more confident in taking charge of discussions and teaching other students. In an advanced major’s class, in which all students have taken the same prerequisites and have had more experiences engaging in discussion, more students may feel confident contributing reasons for their answers or refuting others’ ideas.

Other studies have reported difficulties in training students to recognize features of a good argument. Van Lacum et al. (2014) trained students to follow seven different argumentation characteristics in reading original science literature and tested students both before and after instruction to measure the impact of such training. Students improved in their ability to identify the motive, objective, main conclusion, and implications in the papers they read, but did not improve in their ability to identify supports (warrants) and
counterarguments within the writing. Although students were identifying written versions of argumentation practices rather than engaging in exchanges of ideas, the study highlights that it is challenging for students to learn how to identify and use the building blocks of a high-quality argument.

**Discussion Time**

The students in our study spent a relatively short time discussing their ideas, ~1 min on average, unless LAs were present, in which case they spent ~1.5 min. These times are similar to the length of time reported in other studies of introductory students answering clicker questions (e.g., introductory physics; Miller et al., 2014). Although we only tracked performance on clicker questions by group, rather than by individual student performance, we did not find any correlation between the time taken in discussion and the correctness of the group in its answer. In contrast, Miller et al. (2014) found that individual students who answered clicker questions incorrectly took longer to record their answer, both before and after peer discussion. However, these authors did not measure the length of actual peer discussion, only each individual’s time to response after clicker polling was opened.

Introductory biology students could have relatively short discussions for many possible reasons. Students may not be aware of their lack of understanding (lack of metacognition) and thus may not engage in lengthy discussion of their ideas. They may not feel comfortable challenging incorrect peer ideas or raising alternative ideas. In addition, they may not discuss their answers for a longer time, because they are not concerned about whether they get the question correct or incorrect (in this study, clicker points were for participation only). Assured of participation credit, students might choose to hear the class discussion or instructor explanation of the question rather than engaging with their peers. On the other hand, previous research has shown that awarding points for correctness of groups who engaged in discussion with LAs about half the time, if the question was a lower-order Bloom’s level, students were significantly more likely to articulate reasoning and a single measure of immediate question correctness does not imply that the discussion is not helping students learn. Because the clicker-question responses are not graded, students may feel free to discuss and vote without concern for whether they are correct. Many other factors likely influence whether students fully process the ideas discussed. Ultimately, determining whether discussions involving high-quality reasoning and exchange of questions result in improved undergraduate student performance will likely require using multiple different and longer-term measures of assessment in addition to the immediate in-class diagnostic.

**Cognitive Difficulty**

As previously shown (Knight et al., 2013), the Bloom’s level of a question does not necessarily predict the characteristics of student discussion. However, when looking at the subset of groups who engaged in discussion with LAs about half the time, if the question was a lower-order Bloom’s level, students were significantly more likely to articulate reasoning for multiple answers. This is somewhat surprising, in that we would have assumed that the more cognitively demanding questions would generate more discussion of different answers. However, introductory students may find it easier to discuss a question that experts consider fact based, or even easy, because they may feel confident in helping one another clarify factual information required to answer the question. Thus, they may engage in an exchange of information about the likelihood of each answer as they try to figure out what each answer means. Cognitively challenging questions may also be more difficult for groups to initiate or sustain a discussion around, as the argument needed to support an answer would be more complex. The implication of this finding is that in courses emphasizing cognitively challenging questions, students may require more explicit instruction on how to engage in argumentation.
Group Size
In the discussions we recorded, an average of 3.3 students participated, and the number of speakers across all discussions ranged from one to four speakers, with an occasional discussion involving five. When considering all discussions, a higher number of speakers involved in discussion was positively correlated with all global discussion characteristics (Table 5). When considering a subset of the discussions (only the groups who had a chance to interact with an LA; Table 6), the number of speakers predicted only higher use of requesting information and requesting reasoning but did not have a significant impact on any other characteristics. These patterns were found in an auditorium-style lecture hall, indicating that seating arrangement is not necessarily a limitation to productive discussion among larger groups. This finding aligns with previous studies suggesting that groups should be between three and five to maximize student participation and exchange of ideas (Beichner and Saul, 2003). With more speakers in a conversation, it is more likely that different ideas will be expressed, more questions asked, and potentially higher level exchanges of reasoning. Groups of more than five may be too big to be productive, but we were not able to measure discussions among larger groups. This finding does suggest that one course dynamic component to consider should be the size of groups, with four to five students potentially being better than two or three for stimulating exchanges of ideas.

Study Day
We did not necessarily expect student discussions to improve in quality of reasoning or exhibit more of the other coded characteristics over the short recording period (2.5 wk). However, some topics are much more challenging for students than others, and this could be reflected in the students’ discussions of clicker questions asked on different topics. Thus, we included study day as a factor in the regression analyses and found that this factor was a significant predictor of several discussion characteristics, sometimes inversely. In none of these cases were the odds ratios or betas very high, indicating a limited effect of this variable (Tables 5 and 6). We suggest that the characteristics of student discussions are likely to vary substantially from day to day due to the nature and perceived difficulty of the content being discussed in that class period. Because we can hold this factor constant in the regression analyses while exploring whether other factors impact each discussion characteristic, it does not confound our other findings. However, this finding does support our initial rationale for collecting multiple data points for individual groups of students, as their discussions are different on the different topics. It is also possible that student discussions change substantially over the course of an entire semester: students may better learn how to better interact with one another and may take more or less of an interest in their discussions. To measure this potential, one would need to examine the difference between discussions students have early in a semester (weeks 2–3) to the discussions the same students have much later in a semester (weeks 10–12).

Impact of LAs
At the beginning of our study, we predicted that LA interactions would enhance global reasoning-related discussion features such as Exchange of Quality Reasoning. Instead, we found that LA presence increased student use of reasoning, but not significantly (Table 6). On the other hand, LA presence did impact student use of questioning, shifting them away from using information questions such as “What did you vote?” to questions that were reasoning related (i.e., reasoning statements followed by a feedback request). This increased use of feedback questions is not surprising, as the simple presence of an LA who has previously taken the course is likely to prompt questions from the students. Students also spent significantly more time in discussion when an LA was present; this difference is specific to when the LA is actually interacting with students (Table 6) rather than being a function of an individual group’s dynamics (Table 5). Thus, having an LA engage in discussion with students has the potential to help them discuss the material more thoroughly. However, it is also clear that LAs interact with students in heterogeneous ways, likely accounting for the lack of a significant effect on the Exchange of Quality Reasoning in discussion.

The LAs involved in this course are generally between one semester to 2 yr more advanced than the students enrolled in the course. Thus, they are by no means content experts or experts at teaching, which many studies suggest requires extensive training (e.g., Spillane, 1999; Windshitl, 2004). Perhaps not surprisingly, then, despite pedagogical instruction, LAs do not perfectly execute their interactions with students. In our sample, LAs explained answers to students, using reasoning, in 50% of their discussions. LAs likely think they are helping students build understanding by explaining an answer, when, in fact, the data indicate that providing reasoning often ended the students’ attempts to grapple with the material. Importantly, however, LAs can have a positive effect on student interactions when they use practices that have been demonstrated to draw students into deeper cognitive processing. When LAs used prompting questions, either to encourage student thinking or to more explicitly ask for their reasoning, students readily engaged in these behaviors. A recent study on chemistry peer tutors and their interactions with students during process-oriented guided inquiry learning activities (Kulatunga and Lewis, 2013) showed markedly similar outcomes. In their study, the authors identified a suite of interactions that peer tutors used with students, most of which overlap with our classifications. These authors found that “direct teaching” (our Providing Reasoning category) was much less successful at getting students to generate warrants than “probing and clarifying” interactions (our Prompting Question and Requesting Reasoning categories). We strongly support their conclusion that, when peer tutors combine questioning and mediating behaviors, they are creating an instructional scaffolding that helps students produce high-quality reasoning and deeper understanding and that serves to generate more student questions. All of these interactions are likely to promote the use of reasoning, and help students learn from their in-class discussions.

Caveats
In this study, we collected data from discussions among a relatively small group of students in a single course. It is possible that other students in different courses or even other students in the same context might behave differently. However, the large number of discussions analyzed, and the
relative similarity of the data recorded from these students to data from previously studied advanced students (Knight et al., 2013) and other groups of introductory biology students in different courses and years (unpublished data), indicates that these students are likely to be representative of other students in other courses under similar conditions. The strength of this study is in the number of discussions analyzed, as we can generalize the discussion trends and characteristic interactions for the students who were observed. The coding scheme we have developed and the variables that impact discussion can be used by others to document their own students’ interactions and/or test for additional factors we did not measure.

**Instructional Implications**

While components of argumentation may be part of student thinking processes, they are not often articulated fully during in-class discussions. On the other hand, students are willing and able to engage in argumentation if given the tools and time to do so. If instructors agree there are benefits to engaging students in reasoning during class, they will want to make this connection explicit in their learning objectives and in establishing classroom norms. To increase the use of reasoning in the classroom, instructors can use more specific cues to trigger student articulation of reasoning and can train students in how to construct an argument, taking class time to show students the utility of this practice.

Relatively small changes to cues and practices have the potential to make a difference. For example, instructors could rotate a variety of meaningful cues before each clicker question or longer group discussion that clearly state how the students should interact (e.g., “Explain your reasoning for your answer to your neighbor, and ask for their reason even if you agree on the answer,” or “With your neighbor, explain your reasons, and determine whether you have all the information you need to be confident of your answer,” or “Find someone who chose a different answer and discuss your reasons, even if you have to ask other groups,” or “Don’t be hesitant to challenge someone’s reason in a constructive way”). Subtle changes to what students are asked to do will keep the cues from sounding repetitive and yet continue encouraging students to practice articulating their reasoning.

In addition, seeing what scientists consider to be adequate evidence for drawing conclusions and how they convince one another of the mechanism behind a particular concept are likely essential for helping students understand how to construct their own arguments. This could be achieved in a classroom setting by showing examples of scientists constructing arguments, and then conducting a class-wide discussion of a similar question and its answers using an argumentation framework (such as Toulmin’s), guided by the instructor. This type of modeling should be repeated several times during a course, so the analysis and practice of reasoning becomes a classroom norm.

Perhaps the most important implication of this study is that instructors must not rely solely on general pedagogical training of peer coaches to promote the types of interactions we expect such coaches to encourage. Using prompting or leading questions along with direct requests for reasoning may not be easy for peer coaches to implement in real class settings, even if they have learned about these practices in their training. Showing peer coaches demonstrating that their impact can be positive or negative and that student behaviors are predicted by how they choose to engage with students may remind them to monitor their interactions. In addition, peer coaches would undoubtedly benefit from more hands-on training, with particular attention paid to practicing how to scaffold learning with their students (Kulatunga and Lewis, 2013) and using role-playing strategies to practice encouraging exchange of reasoning. The recommendations made above for instructor–student interactions can be applied to better training peer coaches as well; after all, these coaches are still students and likely need more practice in learning how to use reasoning themselves before they can help others do so. Finally, instructors and peer coaches alike would benefit from working together to explore ways to generate a classroom environment that focuses on deep understanding of content through the explicit use of argumentation and reasoning.

**ACKNOWLEDGMENTS**

This work was carried out with support from the National Science Foundation (DUE 1140789). We are grateful to the members of our grant advisory board: Noah Finkelstein, Derek Briggs, Valerie Otero, Melissa Dancy, and Laurel Hartley and our external evaluator, Sam McKagan, for excellent suggestions and assistance during this project. We are also indebted to Adam Bohr and Roddy Theobald for help with statistical analysis and to the Discipline Based Education Research group of the University of Colorado for ongoing feedback. For help with data collection and transcription, thanks to undergraduate assistants Amedee Martella, Alex Merritt, Nick Myers, Erika Lai, Francis Li, and Sarah Zimmermann. This work could not have been completed without the gracious support of instructors Jennifer Martin and Nancy Guild, and the willingness of University of Colorado students to participate in this research.

**REFERENCES**

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


Redesigning undergraduate biology courses to integrate quantitative reasoning and skill development is critical to prepare students for careers in modern medicine and scientific research. In this paper, we report on the development, implementation, and assessment of stand-alone modules that integrate quantitative reasoning into introductory biology courses. Modules are designed to improve skills in quantitative numeracy, interpreting data sets using visual tools, and making inferences about biological phenomena using mathematical/statistical models. We also examine demographic/background data that predict student improvement in these skills through exposure to these modules. We carried out pre/postassessment tests across four semesters and used student interviews in one semester to examine how students at different levels approached quantitative problems. We found that students improved in all skills in most semesters, although there was variation in the degree of improvement among skills from semester to semester. One demographic variable, transfer status, stood out as a major predictor of the degree to which students improved (transfer students achieved much lower gains every semester, despite the fact that pretest scores in each focus area were similar between transfer and nontransfer students). We propose that increased exposure to quantitative skill development in biology courses is effective at building competency in quantitative reasoning.

INTRODUCTION
The field of biology has become increasingly reliant on interdisciplinary approaches to address complex biological problems, and these approaches typically require significant use of quantitative analysis. However, preparation of students for this new era of quantitative biology has been lacking, in part because undergraduate biology courses have been slow to incorporate quantitative reasoning into the classroom setting (Bialek and Botstein, 2004; Feser et al., 2013). Greater emphasis on quantitative skill development has been noted as increasingly important for preparing biology students for medical (see Association of American Medical Colleges/Howard Hughes Medical Institute report [AAMC/HHMI, 2009]) and graduate school (Barraquand et al., 2014). In addition, lack of training in quantitative sciences has been noted as an impediment to advances in research by practicing biologists (Chitnis and Smith, 2012; Fawcett and Higginson, 2012; Fernandes, 2012). Many recent papers and reports have called for a “revolution” in undergraduate biology education in which analysis, graphical thinking, and quantitative skills receive greater emphasis (e.g., National Research Council [NRC], 2003, 2009; Bialek and Botstein, 2004; Feser et al., 2013; Aikens and Dolan, 2014). A vision of how new curricula might look is laid out in a widely cited American Association for the Advancement of Science (AAAS) report...
Of note, the second core competency listed as necessary particularly in courses taken by first-year students (NRC, 2003). This compartmentalization may give students the false impression that biological knowledge can be acquired with minimal understanding of the more quantitative fields (e.g., mathematics and statistics) and leads to all-too-familiar situations in which students are unable to apply, or transfer, concepts learned in one domain to another, such as applying the math they know in a biological context (NRC, 2000; Haskell, 2001; Mastascusa et al., 2011). Further, the level of quantitative rigor in mathematics and physics courses taught for biology students is often lower than that taught to mathematics, chemistry, physics, and engineering undergraduates (Bialek and Botstein, 2004), which leaves many biology students unable to “think in math,” even if they understand its importance for their field (NRC, 2003). While students often receive some practice in quantitative thinking and analysis in upper-level courses, this may be too late in the curriculum to have much effect. The NRC’s BIO2010 paper specifically encourages integrating mathematics into biology courses early in the curriculum, particularly in courses taken by first-year students (NRC, 2003).

In this paper, we report on the development and assessment of four stand-alone modules that require students to integrate quantitative thinking into an introductory undergraduate biology course focused on ecology and evolution. These modules are designed for use in weekly group-based active-learning class sessions. Each module is designed as a dry laboratory exercise (pencil and paper/computer based) that applies quantitative thinking to biological problems that are ideally covered concurrently in the lecture component of the course. Applying mathematical and statistical concepts in a biological context would, we hoped, teach students that quantitative thinking was a useful, practical, and indeed indispensable way to approach biological problems. We assessed the influence of these modules on development of student competencies in three focal areas, as outlined in the AAMC/HHMI Report (2009): 1) their ability to demonstrate quantitative numeracy and facility with the language of mathematics (skill E1.1 in AAMC/HHMI, 2009, hereafter referred to as “quantitative numeracy”), 2) their ability to interpret data sets and communicate those interpretations using visual and other appropriate tools (skill E1.2, hereafter referred to as “data interpretation”), and 3) their ability to make inferences about natural phenomena using mathematical models (skill E1.5, hereafter referred to as “mathematical modeling”). We assessed these skills with a pre- and postassessment test over each of four semesters in large introductory biology courses (class sizes ranged from ~65 students in the Summer to >250 in the Fall and Spring semesters). We analyzed student gains in reference to demographic and other background student information (e.g., transfer vs. non-transfer student) after the posttest was administered.

We also conducted a small, qualitative, interview-based study of students’ reasoning on quantitative problems. While this study was quite limited in its scope and aims, it provides some insights into the various forms of reasoning students bring to bear on problems that call for quantitative reasoning and highlights certain concepts and terminology of quantitative biology with which students struggle.

**METHODS**

**Module Development**

The first-year biology sequence at University of Maryland–Baltimore County (UMBC) comprises two courses: 1) molecular/cellular biology and 2) ecology and evolution. This paper focuses only on modules used in the eco/evo course. However, we have developed an additional 10 quantitative modules that have yet to be implemented, including four additional modules for use in ecology and evolution and six in cellular and molecular biology. These, and the four modules reported on in this study, are available by request at http://nexus.umbc.edu.

To develop each module, we first chose a topic in the course that we felt naturally lent itself to quantitative treatment and for which we thought quantitative treatment could enhance student understanding of the biological concept. We then designed the modules with several components to make it easy for faculty to adopt them in their courses: 1) a tutor guide that includes an introduction to the biological content; 2) a summary of the contents, including a table aligning each activity with learning objectives and quantitative competencies related to quantitative reasoning (as outlined by the AAMC/ HHMI report, 2009); 3) a list of the mathematical and statistical concepts covered and quantitative skills required and estimated time to complete the in-class component of the module; 4) an in-class worksheet component of the module that the students receive, which provides an introduction to the biological problem and the activities to be completed; 5) preclass exercises designed to review the mathematical and statistical concepts needed to successfully complete the module; 6) a student survey form to be completed online after completion of the module, to obtain feedback on how helpful the module was to their learning; 7) a suggested list of formative assessment questions along with their alignment to AAMC/HHMI core competencies and learning objectives; and 8) a teacher’s guide for implementation. An example of the instructional guide typical of all modules, including a table of learning goals cross-referenced to specific activities in the module, is provided in Appendix 1 in the Supplemental Material.

**Module Implementation**

The in-class worksheet is designed for groups of three or four students to complete in a 50- to 60-min discussion session. At UMBC, these modules were implemented in a smart classroom with a capacity of 90 students in 75-min discussion sections (allowing time to discuss the modules before and after completion). One to two graduate teaching assistants
Quantitative Biology Modules

Vol. 15, Highlights of 2016 77

and two undergraduate teaching assistants were assigned to each section. Course instructors met weekly with the teaching assistants to go over the module activities for the upcoming week and to discuss ways to facilitate student learning in the discussion sessions. These meetings were also used to discuss how well the students did on different aspects of the modules after they had been implemented the prior week and to field suggestions on improving module activities/clarifying questions for the future.

Each group of students in the class had access to computer terminals, although not all modules require their use. Students were asked to complete the premodule exercises before attending the discussion session, and also to carefully read the module, which was enforced with three-question reading quizzes administered either online before class or at the start of each session. Each student submitted a completed worksheet at the end of the discussion session. The modules constituted 20% of the overall course grade.

**Contents of the Four Modules**

A summary of the contents of each module is provided below in the order in which the modules were delivered each semester.

**Mendelian Genetics.** This module allows students to calculate and predict the genotype and phenotype frequencies of a Mendelian trait resulting from monohybrid and dihybrid crosses. Students also learn how to calculate and interpret the results of a chi-squared statistical analysis to test hypotheses about the independent assortment of traits. This module addresses the following general learning goals from the AAMC/HHMI report: students will demonstrate quantitative numeracy (skill E1.1) and make statistical inferences from data sets. The module includes application of basic mathematics to Mendelian patterns of inheritance, including simple calculations of proportions (frequencies), calculations of chi-squared statistics, and a single question on probability. Students also use their data to calculate and interpret the results of a chi-squared statistical analysis to test hypotheses about independent assortment of traits.

**Introduction to Mathematical Modeling.** This module is designed to introduce students to a simple linear mathematical model in the context of negative frequency-dependent selection. This module addresses the following general learning goals from the AAMC/HHMI report: students will demonstrate quantitative numeracy (skill E1.1), interpret data sets, and communicate those interpretations using visual tools (skill E1.2); make statistical inferences from data sets (evaluating best-fit linear relationships based on calculating error sums of squares); and make inferences about natural phenomena using mathematical models (skill E1.5). The module includes data on the frequency of red-finned cichlid fish within a population and the number of offspring produced by red-finned adults as a function of their frequency in the population. The students are asked to graph the data by hand and in Excel, draw a line of best fit through the points on the handwritten graph (by eye), and provide an algebraic formula for this best-fit line. They are then walked through the basics of regression to demonstrate how the best-fit line is determined statistically, by finding the line through the data that minimizes the sum of the squared errors between the actual data points and the line (this is done using calculations in Excel). Students are then asked to interpret the relationship between variables. They are also asked how their interpretation might differ if certain characteristics such as slope and intercept of the line changed. The prelaboratory exercises require students to create a mathematical model in a non-scientific application. For example, one problem states, “On her way to her volleyball game, Joanne stops by the grocery store for a healthy snack that will give her energy for the game. She decides to get a jar of peanut butter and some bananas. She notices that the peanut butter costs $3.99 for a jar and that the bananas cost $0.49 per pound of bananas. 1. How much will one pound of bananas and a jar of peanut butter cost? 2. Joanne decides to bring a snack for each of the 6 girls on the volleyball team. How much will it cost for six pounds of bananas and a jar of peanut butter? 3. On the grid provided below, draw a Cartesian coordinate system with number of pounds of bananas on the x-axis and the total cost of the peanut butter and bananas on the y-axis. Plot the two points that you found for the previous two questions. Plot the line on the grid above through the two points that you found. Now compute the slope of the line that goes through these two points (remember, “rise over run”: (y2 – y1) = m(x2 – x1), where m is the slope of the line). Now write the equation of the line in slope-intercept form, y = mx + b, where b is the y-intercept (the place where the line crosses the y-axis). Congratulations! You have just made a mathematical model of the cost you will pay to provide peanut butter and bananas to your teammates!” The idea is to allow students to practice the type of mathematical thinking they will be using in the module, but to use examples more relevant to their daily lives. After this, the students are given a more biologically oriented example in the prelab using data on conifer density that students are asked to graph and draw a best-fit line through. Similar questions regarding characteristics of the line are asked as in the worksheet itself.

**Population Genetics I—Breeding Bunnies and Natural Selection.** This module explores the effect of natural selection on allele frequencies, and it also allows students to use calculations of expected and observed allele frequencies to determine whether populations are in Hardy-Weinberg equilibrium. This module addresses the following general learning goals from the AAMC/HHMI report: students will demonstrate quantitative numeracy (skill E1.1), interpret data sets, and communicate those interpretations using visual tools (skill E1.2); make statistical inferences from data sets; and explain how evolutionary mechanisms contribute to change in gene frequencies in populations. This module uses as an example a hypothetical recessive trait (furlessness) in a small population of wild rabbits. A collection of red and white kidney beans are used to simulate the genotypes in the population of rabbits, with the red beans representing the alleles conferring the dominant trait (fur), and the white beans representing the recessive alleles (homozygotes are furless). The rabbits are bred by randomly picking two alleles (or beans). The number of rabbits with the recessive allele is recorded (to estimate allele frequency), and the rabbits with two recessive alleles (two white beans) are discarded (reflecting strong truncating selection against these rabbits). The students are asked to repeat the above steps, imposing selection across several generations, and record and plot their data.
(allele and genotype frequencies). The students then draw conclusions regarding the Hardy-Weinberg principle based on allele and genotype frequencies in the data they collected. Finally, the students perform a chi-squared test on the data to test the null hypothesis that the population experiencing selection is in Hardy-Weinberg equilibrium. The prelab exercises present a similar experiment with gumballs, and the students are asked basic probability questions.

**Population Genetics II—Gene Flow and Genetic Drift.** This module is designed to immediately follow the previous module on natural selection of furless rabbits. The learning goals of this module are the same as the other population genetics module discussed above. In this module, again using the furless rabbits example, students explore how gene flow and genetic drift can act to influence the response to natural selection. Two groups of three students each work together, forming a population of rabbits on an island, separated by a river. One side of the river has a mild climate favorable to furless rabbits, and the other side of the island has a harsher climate less favorable to furless rabbits. The students go through the natural selection process described in the previous module for several generations, then the groups exchange three individuals from each population, representing gene flow between the populations on both sides of the river. The students then go through the natural selection process and record the allele frequency in each population. Once the data have been collected, students are asked to draw conclusions about the effect of gene flow and genetic drift on natural selection.

**Assessment Tool**

We developed a pre/posttest for summative assessment of competency in the focal quantitative skills and reasoning. Beginning in the Fall of 2011 and continuing every semester thereafter until Spring 2013, we calculated the discrimination and difficulty of pre-post questions and modified questions to develop a valid pre/postassessments exam. For discrimination, we wanted to determine whether questions were discriminating between high-performing students and low-performing students. Items that had a point-biserial correlation coefficient (Kornbrot, 2014) $<0.20$ were classified as “low discrimination” and were revised for subsequent semesters. A point-biserial correlation coefficient measures the correlation between a continuous variable (in our case, the student score on the posttest) and an ordinal or nominal variable (in our case, whether the student answered an individual question correctly or not). Higher coefficients indicate both a strong correlation between score on the assessment and answering an item correctly and a high level of discrimination.

For difficulty, items that were answered correctly by $\geq 90\%$ of the students who took the exam were deemed too easy, and items answered correctly by $<20\%$ of students were deemed too difficult. These items were also revised before the first formal use of this instrument to obtain data for this study (Spring 2013). The validated pre/posttest was given across four semesters, Spring 2013, Summer 2013, Fall 2013, and Spring 2014 (Appendix 2 in the Supplemental Material). We continued the discrimination and difficulty analyses of the questions on the pre/postexam each semester and made minor changes in two questions (changed distractors) after the Summer of 2013. In addition, two questions did not appear on all exams every semester (see Appendix 2 in the Supplemental Material for details).

The actual content of the test consisted of questions to assess the level at which students have achieved the desired competencies, one attitude assessment question, and other questions about concepts in the course that were not quantitative in nature.

The postassessment exam was given on the final day of discussion. Student motivation was stimulated in two ways. First, we told students that some of the questions or alternative versions of them would appear on the final exam and so this exam was good practice. Second, students were awarded one participation point for a valid attempt at the exam (the exams were curated after they were given, and there was no evidence of any inauthentic attempts [e.g., filling in the same letter for every answer or leaving half the questions blank]).

To see whether individual characteristics of students were associated with changes in student performance, we also asked a series of questions designed to gather demographic and educational history data for each student (Appendix 1 in the Supplemental Material). These demographic surveys were administered immediately after students took the posttest exam in Spring 2013, Summer 2013, and Spring 2014 to minimize the influence of stereotype threat on exam performance (Steele and Aronson, 1995; Spencer et al., 1999).

Further, a qualitative study involving individual interviews with seven students in the freshman-level ecology and evolution class was undertaken in Spring 2015 to investigate how the interviewees understood the pre/posttest questions and what forms of reasoning they used in answering them. We received institutional review board approval for the entire study (UMBC IRB Y13WL04051).

**Statistical Analyses**

We used a nested logistic regression analysis (using SAS, version 9.3) to test for changes in performance in the skills of quantitative numeracy (skill E1.1), data interpretation (skill E1.2), and mathematical modeling (skill E1.5) based on pre-post questions associated with each skill. We used the statistical model $y \leftarrow \text{pre} + \text{post} + \text{pre-post} + \text{subject}$, where $y = 0$ if the question was answered incorrectly, or 1 if correct; $\text{pre} = \text{pre- or posttest}; \text{post} = \text{multiple-choice question nested within either the pre-or posttest};$ and $e$ is the error term. Instead of analyzing the total test score, we partitioned the quantitative questions into three subgroups, with one set of questions for each quantitative skill (to see how questions were partitioned among skills, see Appendix 2 in the Supplemental Material). We then analyzed the data separately for each of the three subgroups, with one analysis for each skill each semester.

We used Cohen’s $d$ to estimate the magnitude of change in the proportion that answered correctly in each skill from the pre-exam to the postexam (Cohen, 1992). Because multiple questions in the pre/postexam addressed each skill, we used the proportion correct for each question associated with a particular skill to calculate Cohen’s $d$. We used the formula $\left( M_e - M_p \right) / SD_{pool}$, where $M_e$ was the mean proportion of correct answers on questions addressing a particular skill in the posttest, and $M_p$ was the mean proportion of correct answers on questions addressing that skill in the pretest.
Student Interviews

Individual interviews were conducted with seven undergraduate students enrolled in the ecology and evolution course to inquire into their reasoning processes in answering four questions (questions 2–5) excerpted from the pre/posttest (see Appendix 2 in the Supplemental Material). The interviews were conducted during the 11th and 12th weeks of the semester, after students had already participated in a subset of the quantitative modules (Appendix 4 in the Supplemental Material). The purpose of the interviews was to provide data about the nature and quality of students’ reasoning on questions related to the quantitative thinking portions of the modules. Our research questions for this portion of the study were

1. What types of reasoning did students use to answer these questions?
2. For questions that students answered incorrectly, which forms of reasoning were they using and which aspects of the concepts did they struggle with?
3. Was there a discernible pattern in reasoning processes among students who answered the same questions incorrectly?
4. Was there a discernible pattern in reasoning processes among students who performed similarly in terms of their final course grades?

Students were selected for interviews by their instructor based on their average grades on the first two exams. (The interviewers did not know the performance level of any of the students at the time of the interview.) Initially, nine students were chosen to participate in the interview process (three who had “A’s” at the time of the interviews, three who had “B’s,” and three who had “C’s,” all of whom were chosen randomly from within each letter grade). Of the nine who originally agreed to the interview, only seven actually participated (three with “A’s,” two with “B’s,” and two with “C’s” at the time of the interview). Each interview lasted ~20 min and involved an interviewer engaging each student in a semistructured review of their answers to all four test questions after they had read and answered them. The interview questions were designed to elicit and probe the students’ thought processes by having them explain how they came to their answers. (See the Interview Protocol in Appendix 3 in the Supplemental Material.) All of the interviews were audio recorded, and an observer listened and took notes. The interviews were transcribed from the observer’s notes and the audio recordings.

A grounded theory approach (Corbin and Strauss, 2014) was used to code the interview transcripts for the types of reasoning students used in their explanation of their answers for the test questions. This approach allowed us to derive the categories of reasoning the students were using directly from their utterances, rather than to predict and apply a priori certain forms of reasoning to our analysis of the interview data. Such a methodological stance is common in qualitative research, in which the experiences of the participants in the research are considered to be fundamental to the development of theory.

RESULTS

Assessment of Student Learning Outcomes from Pre/Postexam

The first question on the pre/postexamination asked students how useful they thought quantitative approaches (e.g., mathematical modeling, statistical analyses) are to the study of biology. In all four semesters, students came in with a fairly strong opinion that quantitative approaches were important for studies in biology (Table 1). In every semester, greater than 90% of the students thought that quantitative approaches were, at a minimum, “very important” for studying modern biological problems (summing the top three rows of Table 1). This attitude was little changed (a 1% or less increase) at semester’s end.

The remaining portion of the pre/postexamination was designed to assess students in the three focus skills described above: quantitative numeracy, data interpretation, and mathematical modeling, aligning with AAMC/HHMI (2009) skills E1.1, E1.2, and E1.5, respectively. While these skills are not mutually exclusive, an effort was made to test each skill separately in the pre/postexam (see Appendix 2 in the Supplemental Material for categorization of questions). All modules required students to use quantitative numeracy and data interpretation skills; however, only one explicitly required students to develop, use, and interpret mathematical models. Although there was slight variation from semester to semester (two of the validated questions did not appear each semester), the pre/postexam used approximately five questions to assess quantitative numeracy, six questions to assess data interpretation, and three questions to assess mathematical modeling.

Proficiency of Students in Quantitative Numeracy

Students improved their quantitative numeracy from the pre- to posttest, although the magnitude of the effect varied...
The Ability of Students to Make Inferences from Mathematical Models in Biology

Gains in the ability of students to make inferences about natural phenomena using mathematical models were highly variable across semesters, but on average were slightly lower than in the other two skills (Table 2 and Figure 1C). As was the case for quantitative numeracy, there was a broad range in pretest scores from semester to semester (range: 51% in Fall 2013 to 65% in Summer 2013), but unlike the results for quantitative numeracy, there was also a broad range of scores across semesters in posttest results (from 58% in Fall 2013 to 69% in Spring 2013).

Student Demographics and Background Characteristics on Gains

We gathered demographic and background data on students to examine their potential influence on the posttest scores in the different quantitative skills. In the three semesters for which we had demographic data, just more than half the students were female. Greater than 70% were either white or Asian. In Spring 2013 and Spring 2014, most students had a grade point average (GPA) of 3.1–4.0 (65 and 59%, respectively). Fewer students had a GPA of 3.1–4.0 in Summer 2013 (48%). Students in Summer 2013 were less likely to have completed math and physics courses in high school than students in either Spring 2013 or Spring 2014. They were also less likely to have taken calculus 1 or 2 at UMBC. Summer 2013 had the most transfer students (33% compared with 16% in Spring 2013 and 27% in Spring 2014).

Table 2. Results from the nested logistic regression analysis of student performance in pre/postexam scores (p values are those associated with the overall treatment effect in the model) in quantitative numeracy, data interpretation, and mathematical modeling

<table>
<thead>
<tr>
<th>Skill</th>
<th>Spring 2013</th>
<th>Summer 2013</th>
<th>Fall 2013</th>
<th>Spring 2014</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p</td>
<td>d</td>
<td>p</td>
<td>d</td>
</tr>
<tr>
<td>Quantitative numeracy</td>
<td>&lt;0.0001</td>
<td>0.4</td>
<td>0.03</td>
<td>0.2</td>
</tr>
<tr>
<td>Data interpretation</td>
<td>0.001</td>
<td>0.4</td>
<td>0.01</td>
<td>0.1</td>
</tr>
<tr>
<td>Mathematical modeling</td>
<td>0.001</td>
<td>0.9</td>
<td>NS</td>
<td>–</td>
</tr>
</tbody>
</table>

*p Values reported are for Wald’s chi-squared values. d Values are Cohen’s d effect sizes. All significant changes from the pre- to posttest scores represent positive gains in student skills.
We used multiple regression analysis to determine whether any of these demographic or student background variables were significant predictors of student scores in the posttest exam. We included the pretest score of each student in this model as well (as suggested by Theobald and Freeman, 2014). Results are shown in Appendix 5 (Tables A–C) in the Supplemental Material. For quantitative numeracy (Appendix 5A), in Spring 2013, three characteristics were significant predictors of postassessment performance: the pretest score, student transfer status, and whether the students had taken precalculus in high school. On the basis of the regression coefficients of each significant effect for each additional point a student scores on the preassessment, we expect a student’s score on the postassessment to increase by 0.273 points ($\beta = 0.273, p < 0.0001$). Transfer students were expected to score 8.072 points lower on the postassessment than nontransfer students ($p < 0.05$). Students who took precalculus in high school were expected to score 9.853 points lower on the postassessment than students who did not take precalculus in high school ($p < 0.05$). In the Summer of 2013, only the preassessment score was related to postassessment performance in quantitative numeracy; a student’s score on the postassessment is expected to increase by 0.444 points for each additional point scored on the preassessment ($p < 0.01$). In Spring 2014, four variables predicted student scores on the posttest: student pretest scores, transfer status, whether students had taken high school algebra, and whether students had AP credit for calculus or were taking calculus one at the same time as they were enrolled in the course. A student’s score on the postassessment was expected to increase by 0.384 points for each additional point the student scored on the preassessment ($p < 0.0001$). Transfer students were expected to score 11.493 points lower on the postassessment than nontransfer students ($p < 0.0001$). Students who took algebra 1 in high school were expected to score 5.708 points lower on the postassessment than students who did not take algebra 1 in high school ($p < 0.05$). Students who received AP credit for calculus 1 or who already took or were currently taking calculus 1 at UMBC were expected to score 8.736 points higher on the postassessment than their counterparts ($p < 0.01$).

Results for data interpretation are shown in Appendix 5B in the Supplemental Material. In Spring 2013 and Summer 2013, the only variable related to postassessment performance was the score on the preassessment. In Spring 2013, for each additional point scored on the preassessment, a student’s score on the postassessment was expected to increase by 0.390 points ($p < 0.0001$). In Summer 2013, for each additional point scored on the preassessment, a student’s score on the postassessment was expected to increase by 0.754 points ($p < 0.0001$). In Spring 2014, two variables were significant; the score on the pretest and whether the student had AP credit for calculus or was taking calculus at UMBC at the time. In this case, for each additional point scored on the preassessment, a student’s score on the postassessment was expected to increase by 0.356 points ($p < 0.0001$). In addition, students who received AP credit for calculus 1 or who already were currently taking calculus 1 at UMBC were expected to score 14.520 points higher on the postassessment than their counterparts ($p < 0.0001$). Results for mathematical modeling are shown in Appendix 5C in the Supplemental Material. In Spring 2013, two variables were significant; the pretest score and whether students received AP credit for calculus 1 or already were currently taking calculus 1 at UMBC. For each additional
point scored on the preassessament, a student’s score on the postassessament was expected to increase by 0.257 points ($p < 0.0001$). Students who received AP credit for calculus 1 or who already took or were currently taking calculus 1 at UMBC were expected to score 16.447 points higher on the postassessment than their counterparts ($p < 0.0001$). In Summer 2013, no variables were related to performance on the postassessment for this skill. In Spring 2014, three variables were significant: the score on the preassessment, transfer status, and whether students received AP credit for calculus 1 or had already taken or were currently taking calculus 1 at UMBC. For each additional point scored on the preassessment, a student’s score on the postassessment was expected to increase by 0.240 points ($p < 0.0001$). Transfer students were expected to score 12.648 points lower on the postassessment than nontransfer students ($p < 0.05$). Students who received AP credit for calculus 1 or who had already taken or were currently taking calculus 1 at UMBC were expected to score 12.081 points higher on the postassessment than their counterparts ($p < 0.05$).

Investigation of Students’ Reasoning on Selected Prel/Posttest Questions

Three main categories of reasoning were discerned in these interviews: quantitative, disciplinary (biology), and general logic. To distinguish these categories of reasoning, we read and reread the transcripts several times, looking for instances in which students explicated their reasoning processes and analyzing them for similarities and connections. Initially, we detected six types of reasoning that students applied in their explanations: 1) reasoning in which students referred to the graph provided in the test questions or other graphic knowledge (quantitative reasoning; graphic); 2) quantitative forms of reasoning that did not specifically mention the graph (quantitative reasoning; general); 3) reasoning through use of biology concepts (biology knowledge: conceptual); 4) reasoning through appeal to how research in biology is conducted (biology knowledge: process); 5) rationales based on how tests are constructed and/or how they should be taken (test-taking knowledge); and 6) general, nonquantitative, non–disciplinary-specific knowledge or logic (general knowledge).

While each of these six types of reasoning seemed to be distinct, and each occurred in more than one interview, distinguishing between, for example, quantitative reasoning that specifically referred to the graph and that which made no reference to the graph or graphic knowledge was not relevant to our purposes. Thus, we combined these two categories into one category of quantitative reasoning. Likewise, it was not necessary to distinguish reasoning that appealed to knowledge of biological concepts from that which appealed to knowledge of how biological research was conducted, so those two categories were combined into one type of reasoning we refer to as disciplinary knowledge. And, finally, distinctions between reasoning that appealed to knowledge of best practices in test writing or test taking and other more general forms of logic were not relevant to our purposes, so we combined those two categories into one general logic category.

Of the three main categories of reasoning, by far the one used most prevalently was quantitative reasoning. The fact that quantitative reasoning was used so widely is not surprising, given that the questions each referenced a graph that preceded the question and each of the questions and their answer options referred to quantitative reasoning concepts. For example, one high-performing student reasoned his way to the correct answer on question 5 (see Appendix 2 in the Supplemental Material), saying, “I decided that [the most reasonable hypothesis is] c. [Student reads the answer from the test sheet aloud.] When density is between 20–30 plants, we would expect no more than 20 seeds per plant. [Interviewer: You mean 200?] 200, yeah. Because between 20 and 30 for the plant density, we definitely see that there’s no data points above that, so we shouldn’t expect, if it were to follow this trend, to be like that. And I just decided that the other answers were either too specific or incorrect.” We coded this part of his explanation as quantitative reasoning, because the student refers to information that is provided only in the graph, and he correctly interprets the graph to say that there are no data points above 200 seeds per plant in the plant density range of 20–30 plants.

It is notable, however, that the interviewed students did not always apply quantitative reasoning concepts accurately, and at times, they used other forms of reasoning as well as quantitative reasoning, including approaches commonly taught as test-taking strategies (coded as general logic) and knowledge of how field biologists conduct research (coded as disciplinary knowledge). A lower-performing student answered question 5 correctly also, but his thinking was based less on reasoning around his reading of the graph and more on the wording of the answers and which one sounded, in his words, “most reasonable.” He did use some quantitative reasoning in eliminating answer “a,” noting: “There’s a wide range of seeds but they all look higher than as soon as you have more density, so I guess that’s an effect.” Likewise, to eliminate answer “b,” he looked at the graph, “If you look at the chart, after 30 it’s not the highest. Ten to 20 is a much higher rate of chance than 30–40.” But in choosing answer “c,” his rationale was based more on the way it sounded: “The way that one is worded it just seems to me like it could actually be something that you could use as a hypothesis to apply to something else ...” It’s using this information to explain to you that if you have this range of plants based on this information we shouldn’t see any more than this number of seeds. Because it’s between this range, correlated with this information from this other place.”

He drew on all three types of reasoning—quantitative, logical, and disciplinary—to eliminate answer “e.” Reading from the answer, he said: “When plants get to a density of 50, we definitely see that there’s no data points above 200 seeds per plant in the plant density range of 20–30 plants.

The question that posed the most difficulty for the students interviewed was question 2. Three of the seven interviewees answered this question incorrectly, and each of their incorrect responses was different. One of these students relied on quantitative reasoning to choose answer “e” (40 plants). Although he was able to use the key to focus correctly on the red marks on the graph for Assateague Island, he seemed to have confused the notions of plant density and plant sampling.

Another student who answered this question incorrectly, chose “d” (20 plants), noting in his explanation that he first read the background information, then looked at the chart,
and then looked at the question. He said, “None of [these sources of information] told me a specific number where it said how many populations were selected. So I counted the individual observations and that was my answer … I assume that these [pointing to dots on the graph] are each individual plants.” This student seems to have drawn on general test-taking strategies (an approach that we coded as logic based) at first to see whether the answer was explicitly stated in either the background information or on the graph. When it was not, he then made an assumption (correctly) that each of the points on the graph represents a sample. His error was in not realizing, or perhaps not reading carefully enough to notice, that the question specifically asked for the number of samples of plants on Assateague, not both islands.

The third student who answered question 2 incorrectly selected “a” (100 plants) as her answer. She indicated that she had multiplied each plot point on the graph by the corresponding plant density on the x-axis and then added them together to arrive at an estimate that roughly corresponded to 100. She said: “I kind of tried to add them up, but I didn’t really add them all up. I just gave a rough estimate. I mean, adding up plant density for Assateague Island because it says that plant density is based on counts of the number of individual plants. So I just added up x values of the square ones, and 100 is too large of a number. But it was much more greater than 40, so it must be around 100.” So, although she was using quantitative reasoning, like the other two students who answered this question incorrectly, she seems to have confused plant density with the sampling of plants, not realizing that each plot point represents a sample.

In general, the students who ended the course with higher grades were more likely to use quantitative reasoning as their major or sole source of reasoning in answering the questions. Lower-performing students also used quantitative reasoning, but often poorly or incorrectly, and often supplemented with general logic or disciplinary knowledge when their quantitative reasoning skills were not sufficient to the task of answering the question. For example, lower-performing students chose answers that reversed the causal relationships from what was displayed in the figure, were unable to recognize whether a figure displayed a positive or negative correlation between variables, did not understand what a change in slope of a regression line would represent for the relationship between two variables, and had difficulty understanding the concept of individual points on a figure representing data samples taken at different time points. However, these students sometimes reached the correct answer to a question using appropriate biological logic, for example, “Well, if there are more plants in this area, they must be more crowded, so it doesn’t make sense that each one would make more seeds.”

For some questions, nearly all students struggled to apply quantitative reasoning appropriately, due to lack of familiarity with the quantitative concept or term. For example, their explanations suggested that few of the students interviewed understood what it meant for one variable to be “sensitive” to changes in another (question 4, answers “b” and “c”).

**DISCUSSION**

We saw significant improvement in all skill areas tested in most semesters, though the effect sizes were often modest and varied from semester to semester. On the basis of the average gains in performance, students improved the most in skills reflecting quantitative numeracy. Their ability to make biological inferences from mathematical models and to interpret data in biological contexts (primarily interpreting graphical results) showed smaller gains. To the extent that the module activities contributed to this skill development, we hypothesize that the prelaboratory activities, which reviewed key skills needed before students came to class, and the repeated application of similar skills to different biological settings were important aspects of the modules that contributed to the learning gains of students. We feel that it is also important to allow students to practice using quantitative skills in the lecture portion of the course (in small groups or through clicker questions). This emphasizes to students the importance of quantitative skill development and integrates the lecture and module activity components of the course. Further research is necessary to tease out the contributions of each aspect of the design and implementation of the modules toward enhancing students’ performance.

The fact that students improved most in quantitative numeracy is not surprising, given that they used this skill in each module during the semester. The other skills were used less often (in particular mathematical modeling, which was only formally used in one module), and students may need additional practice to achieve greater gains in these competencies. Of course, the smaller gains could also be attributable to other factors, such as the fact that interpreting data and using mathematical modeling are difficult skills to master, the modules were perhaps not very effective in improving these skills, there may have been poor alignment between modules and assessment, or our assessments may not have been sensitive enough to measure improvements that were made.

The variability in gain also seems to be based on how much scope for improvement students had coming into the course; gains were typically higher in semesters when the pretest scores were lowest. Despite the fact that the course was taught by four different instructors in different semesters, the degree of competency attained in each skill (at least as reflected by the posttest scores) was similar in each semester, regardless of the level of proficiency that students had coming into the course.

The degree of improvement was greater for some populations of students than others. Aside from the predicted case, in which students who did well on the pretest did well on the posttest, the most general finding was that transfer students showed lower gains during the semester in skills associated with quantitative numeracy and mathematical modeling compared with their classmates who were nontransfer students. Many other studies have also noted poorer performance of transfer students (e.g., Graunke and Woosley, 2005; Duggan and Pickering, 2008), especially those majoring in mathematics and the sciences (Cejda et al., 1998; D’Amico et al., 2014). There could be many reasons for this, including poorer academic preparation, differences in demands on their time (e.g., jobs, family obligations), and differences between transfer and nontransfer students in their sense of community on campus (Townley et al., 2013). Understanding the reasons for the disparity of gains between transfer and nontransfer students will require further study. The only other fairly consistent demographic that predicted enhanced student performance was having received AP
credit for calculus or taking calculus concurrently with the introductory biology course in ecology and evolution. None of the modules required calculus, however, and the only pre/corequisite for our course is precalculus. As such, we suspect that the explanation has less to do with preparation in calculus than it does with general student academic preparation. There is also the possibility that prior experience with calculus develops higher-order problem-solving skills. Again, more research is needed in this area to see whether this is a general pattern for other courses (particularly in science, technology, engineering, and mathematics [STEM] areas) and to understand the reasons for this, if so.

There was one curious finding from our demographic analysis related to posttest performance on questions related to quantitative numeracy: students who took algebra 1 in high school scored significantly lower on the postassessments than students who did not take algebra 1 in high school. Although this relationship was only formally significant in one of the semesters (Spring 2014), the sign and magnitude of the regression coefficient was negative and similar in magnitude every semester, suggesting that this is a consistent relationship. While we do not know the exact cause of this relationship, we note that the sign of the regression coefficients associated with students who had taken algebra 2 in high school was fairly large and positive in two of the three semesters measured (indicating a positive influence of completing algebra 2 on quantitative numeracy, although this effect was not significant in our analysis). This would indicate that, if algebra 1 is the highest level of mathematics completed in high school, this has a negative effect on student performance in tasks related to quantitative numeracy as freshman college students.

A number of efforts to infuse more quantitative reasoning into biology courses have been described in the literature, and many of these have seen positive effects on both student attitudes (Matthews et al., 2010; Thompson et al., 2010; Goldstein and Flynn, 2011; Barsoum et al., 2013) and skill development (e.g., Thompson et al., 2010; Colon-Berlinger and Burrowes, 2011; Goldstein and Flynn, 2011; Madlung et al., 2011; Hester et al., 2014). For example, Hester et al. (2014) designed a new course in molecular and cell biology that required greater reliance on mathematical skills and reasoning to understand biological concepts. They found that students initially had poor ability to apply their quantitative skills to biological contexts but improved when quantitative reasoning was integrated throughout the course, including in in-class exercises and on exams. Another course redesign was implemented with development of a new textbook that integrated more quantitative reasoning into introductory biology courses (Barsoum et al., 2013). The new course and accompanying text focused more on the process of science and less on biological content knowledge. Data interpretation by students taking the new course was significantly better than those taking the traditional course. In addition, despite the decrease in content focus, students taking the new course did just as well on tests of biology content knowledge as did those in the traditional course. Thus, the improved quantitative skills did not come at the expense of biological knowledge, a finding consistent with at least one other study (Madlung et al., 2011). Interestingly, the difference in quantitative skills between students in the new and traditional course disappeared when students were tested on similar skills one semester removed from the experience. Speth et al. (2010) found that infusing quantitative concepts using modules throughout an undergraduate biology course, including on the assessments, led to a significant improvement in their students’ ability to interpret graphically represented data and to create graphical representations of numerical data. They did find, however, that students still had difficulty in constructing scientific arguments based on data, and they note that this is a difficult skill to master in only one semester.

In our case, the gains in student skills were fairly modest, which could be explained by two major factors. First, students are often attracted to biology because they perceive it as the least quantitative of STEM disciplines. As a result, they may be less academically (and even emotionally) prepared to integrate quantitative reasoning in biological settings. Second, the changes we imposed occurred in a single course, with only four modules implemented across the semester. Many authors, as noted above, have emphasized that short-term exposure is unlikely to make a lasting difference in students’ skills and attitudes (e.g., Barsoum et al., 2013; van Vliet et al., 2015). The findings of our small qualitative study of students’ reasoning seem to offer further support for this notion. Although all the students we interviewed demonstrated their ability to apply quantitative reasoning to problems that clearly called for it, their application of such reasoning was uneven and at times incorrect, which may suggest that it takes further exposure and practice for these skills to be well integrated into students’ thought processes. Thus, we believe that it is likely that greater and longer-lasting improvements in quantitative competencies will only be possible through broader curricular reform, including consistent integration of mathematics and statistics (ramping up the sophistication of each across the levels) throughout the biology curriculum. Even within a course, stand-alone modules may not be the best way to promote our goals; Hester et al. (2014, p. 55) report, “We found that stand-alone modules, done on the students’ own time, perpetuated the perception that math was an ‘add-on’ that was not representative of the core content in biology.” Some have even argued that biology, mathematics, physics, and chemistry should be taught in a fully integrated manner (Bialek and Botstein, 2004), and there appears to be some movement in this direction (e.g., Depelleau et al., 2010). Because the barriers to this integration are so high and so many, we continue to see value in our approach. Alternatively, our approach may be seen as the first step toward a truly interdisciplinary curriculum for life sciences students.

The question to which we must always return in pedagogy is: What is the very best use of the very limited time we have in the classroom with our students? Fortunately, this is an empirical question. Though still promulgated, the view that employing deeper learning techniques such as our modules requires reduction in the number of topics discussed or results in students attaining lower degrees of content knowledge has been empirically overturned (e.g., Fatmi et al., 2013; Freeman et al., 2013; Nanes, 2014). Courses and curricula organized as if our role is to pour our knowledge into students’ brains, a school of thought that is still prevalent, harm those students by focusing on lower-level cognitive skills to the exclusion of analysis and critical thinking (Momsen et al., 2010).
Our goal should be to equip students with the cognitive tools they need to investigate questions of interest to them (Gross, 2004), and the evidence is clear that students can only acquire such tools via active application of course concepts to realistic and significant problems. To be effective, these pedagogical reforms must escape single courses of individual motivated faculty and become part of the fabric of undergraduate science education.

ACKNOWLEDGMENTS

We thank the many people who helped us with this project. Sarah Hansen helped with data analysis. Archer Larned, Chia hua Lue, Sarah Luttrell, Karan Odom, and Tony Williams implemented the modules and assessments and gave us a great deal of constructive feedback. Brian Dalton, Chia hua Lue, Michael Martin, and Tracy Smith were graduate students who developed much of the content of the modules. We are also grateful to our colleagues Tony Smith, Tamara Mendelson, and Kevin Omland, who supported the use of these modules in their courses. William LaCourse and Kathy Sutphin provided organizational support. We also thank our colleagues from the NEXUS project for advice, especially Kaci Thompson, Edward (Joe) Redish, Michael Gaines, Jane Indorf, Daniel DiResta, and Liliana Draghici. Special thanks to David Hanauer for his patient and incisive guidance on the research component of this project. This project was funded in part by a National Institutes of Health National Institute of General Medical Sciences grant (1T36GM078008) to Lasse Lindahl (UMBC), and an award from the HHMI to the UMBC under the Precollege and Undergraduate Science Training Program. Thanks to Tim Ford for help with the figures. We also thank two anonymous reviewers and Janet Batzli for helpful comments on the manuscript. All modules and course materials are freely available at http://nexus.umbc.edu or can be obtained by contacting the corresponding author.

REFERENCES


Fernandes AD (2012). No evidence that equations cause impedes communication among biologists. Proc Natl Acad Sci USA 109, E305.


Article

Features of Knowledge Building in Biology: Understanding Undergraduate Students’ Ideas about Molecular Mechanisms

Katelyn Southard, Tyler Wince, Shanice Meddleton, and Molly S. Bolger

Department of Molecular and Cellular Biology, University of Arizona, Tucson, AZ 85721

Submitted May 18, 2015; Revised November 6, 2015; Accepted November 16, 2015

Research has suggested that teaching and learning in molecular and cellular biology (MCB) is difficult. We used a new lens to understand undergraduate reasoning about molecular mechanisms: the knowledge-integration approach to conceptual change. Knowledge integration is the dynamic process by which learners acquire new ideas, develop connections between ideas, and reorganize and restructure prior knowledge. Semistructured, clinical think-aloud interviews were conducted with introductory and upper-division MCB students. Interviews included a written conceptual assessment, a concept-mapping activity, and an opportunity to explain the biomechanisms of DNA replication, transcription, and translation. Student reasoning patterns were explored through mixed-method analyses. Results suggested that students must sort mechanistic entities into appropriate mental categories that reflect the nature of MCB mechanisms and that conflations between these categories is common. We also showed how connections between molecular mechanisms and their biological roles are part of building an integrated knowledge network as students develop expertise. We observed differences in the nature of connections between ideas related to different forms of reasoning. Finally, we provide a tentative model for MCB knowledge integration and suggest its implications for undergraduate learning.

INTRODUCTION

Many national calls for instructional reform and decades of evidence-based research on student learning have convinced us that teaching and learning in the field of molecular and cellular biology (MCB) is difficult. The complexities of the mechanisms involved, the rapidly advancing knowledge base, and the specialized language create a daunting environment for learners and a difficult task for instructors. Several features may contribute to difficulties of learning in MCB, such as the “invisible” and “inaccessible” nature of the entities studied (Gilbert et al., 1982; Duncan and Reiser, 2007), the necessity of mapping across physical and ontological levels (Marbach-Ad and Stavy, 2000; Duncan, 2007; Duncan and Tseng, 2011; Van Mil et al., 2013), and the interrelated and overlapping nature of complex phenomena (Duncan, 2007; Duncan and Reiser, 2007; Duncan and Tseng, 2011; Van Mil et al., 2013). However, literature on the nature of knowledge and conceptual change in MCB is less well developed (with notable exceptions, including such work as Venville and Treagust, 1998; Duncan, 2007). Unanswered questions include: How do students draw connections between ideas to build expert knowledge networks? How do they organize ideas when learning about cellular phenomena? What are the underlying mechanisms of conceptual change that may lead to scientifically non-normative ideas and misconceptions? Building from cognitive learning science theories and the literature describing domain-specific reasoning strategies, we explore undergraduate students’ understanding of foundational MCB principles, specifically the processes of DNA replication, transcription, and translation.

We propose that the complex nature of MCB requires learners to build highly integrated knowledge networks that are both productively organized and flexibly dynamic.
to engage in problem solving and explanation building. We hypothesize that learners must integrate pieces of knowledge, manipulate connections between ideas, and build a cohesive view of the relationships between phenomena to truly build understanding of mechanisms in molecular biology. Although much work has been done to characterize the nature of knowledge among novice learners in science (Gobbo and Chi, 1986; diSessa, 1988; Vosniadou and Brewer, 1992; Southerland et al., 2001; Chi, 2013) and experts in science (Chi et al., 1981; Machamer et al., 2000; Darden, 2002, 2008; Craver, 2002), we know much less about how scientific knowledge develops as students move through an undergraduate major.

In this study, we provide a new lens with which to explore undergraduate students’ understanding of concepts within molecular biology courses. By characterizing undergraduate students’ understanding of key molecular principles, using the knowledge-integration approach to conceptual change, we explore the nature of knowledge and the ways in which it is organized by students, productively and unproductively, as they begin to grapple with the complexity of molecular phenomena.

THEORETICAL FRAMEWORK

The Theory of Knowledge Integration

Learning is a dynamic process involving the acquisition of new ideas, the development of connections between ideas, and the reorganization of prior knowledge (Clark and Linn, 2003, 2013). Much research in the learning sciences has focused on conceptual change, the process by which new knowledge is built (Ozdemir and Clark, 2007; Vosniadou, 2013b). The knowledge-integration theory of conceptual change focuses on the ways in which ideas are sorted, connected, and integrated. Knowledge integration is, by definition, the process by which learners not only add new ideas or facts but also sort through connections and move toward developing a cohesive mental model of a phenomenon (Clark and Linn, 2003, 2013). In doing so, one must sort ideas into appropriate categories, and make connections between ideas, thereby integrating ideas to build a network of knowledge. Knowledge integration is also significantly impacted by the amount of instructional time allocated to important topics. Clark and Linn (2003) demonstrated that significant enhancement of knowledge integration around foundational concepts in thermodynamics resulted from increased instructional time in middle school classrooms, thereby underscoring the idea that deeper learning takes time.

A key component of the theory of knowledge integration is that learners must sort through ideas to construct a cohesive view of a phenomenon. Examples of this sorting include the processes of coalescence and differentiation. Coalescence is the process by which two ideas are merged or combined, whereas differentiation involves the splitting of an idea into separate elements (Clark and Linn, 2013). For example, combining the ideas of heating and cooling into one model of thermal equilibrium would be the “coalescence” of two previously separate ideas. Alternatively, the separation of heat and energy from the idea of temperature would be “differentiation.” While knowledge integration, including the active sorting and restructuring of ideas, is necessary for building scientific knowledge, not all actions of restructuring, integration, coalescence, and/or differentiation are immediately productive (Clark and Linn, 2013). That is, as new connections are made, they may reinforce misconceptions or create non-normative scientific explanations. Non-normative explanations are defined as student ideas that would not be generally accepted by the scientific community, though they may represent logical trains of thought on the part of the student. These are ideas that would typically be considered “incorrect” by an instructor. By contrast, a normative explanation reflects current norms of the scientific community and would generally be considered “correct” by an instructor.

Another key component of the theory of knowledge integration is the creation of connections between ideas. For example, when learning about “mutations,” students should form connections between this idea and the concepts of “gene” and “evolution.” Some connections made by students are spontaneous and temporary, whereas others are stable and persistent. Additionally, not all connections give rise to scientifically normative ideas, but they can also give rise to conflicting and dissenting ideas within a conceptual ecology (Clark and Linn, 2013).

Possibly the most complex component of the theory of knowledge integration is the integration of ideas to build complex knowledge structures. According to Clark and Linn (2013), at its most basic level, “integration involves creating or reinforcing the connection between two ideas” (p. 522). Along these same lines, when investigating students’ understanding of science at different levels of knowledge integration, Clark and Linn (2003) defined a “nuanced” level of understanding in which answers “involve not only the normative ideas but also important connections to other normative ideas or evidence” (p. 459). It is also important to consider that the process of integrating knowledge, including restructuring and reorganizing ideas, leads to complex, dynamic, and emergent knowledge networks (Brown, 2013). As researchers, in order to explore how students are integrating ideas, several conceptual change theories agree that “we need to move away from thinking about conceptual changes as involving single units of knowledge to systems of knowledge that consist of complex substructures that may change gradually in different ways” (Vosniadou, 2013a, p. 12).

To study these complex substructures, we must approach knowledge about this topic as a dynamic ecology of ideas, constantly under reconstruction and restructuring.

In discussing how ideas are sorted, connected, and integrated, it is important not only to look at what ideas are connected but to explore the nature of how the ideas are connected. Students’ motivation for creating particular connections, the robustness of those connections, and the nature of the reasoning that forms a connection between two ideas are important features of knowledge integration that, to our knowledge, have been largely unexplored. We propose that connections between ideas in molecular biology are constrained by multilevel, molecular mechanistic reasoning that may guide developing knowledge to reflect the nature of biological phenomena.

The Nature of MCB Knowledge

In MCB, specific forms of reasoning are required to navigate the unique complexities of the domain; this process often
includes employing reasoning strategies that mirror the nature of the phenomena themselves. "Mechanistic reasoning" is one such type of reasoning required to study the interacting molecular mechanisms that underlie biological phenomena in the field of molecular biology (Machamer et al., 2000; Russ et al., 2008). Understanding these complex biological mechanisms involves 1) recognizing the involvement of specific entities (objects), 2) recognizing the activities created by the interactions of the entities, and 3) understanding the temporal and spatial interactions and organization of these components that create the biological phenomena (Machamer et al., 2000; Russ et al., 2008; Van Mil et al., 2013).

For example, thinking about the molecular mechanism of translation (the synthesis of proteins based on a mRNA molecule) requires the recognition of key entities such as "RNA transcript," "ribosome," "tRNA," and "amino acids," among others. It also involves recognizing activities such as "binding" of the tRNA to the RNA transcript and ribosome and "recognition" of the ribosome binding site on the RNA by the ribosome. And finally, an understanding of the dynamic sequence of events and the necessary physical organization of the molecules allows for a big-picture understanding of the phenomenon. Craver (2001) claims that understanding a sequence of events and the necessary physical organization of specific entities (objects), 2) recognizing the activities created by the interactions of the entities, and 3) understanding the temporal and spatial interactions and organization of these components that create the biological phenomena. (Machamer et al., 2000; Russ et al., 2008; Van Mil et al., 2013).

Molecular Mechanisms

We selected three molecular mechanisms to explore knowledge integration in undergraduate MCB: DNA replication, transcription, and translation. These three molecular mechanisms were selected based on their foundational importance to biological phenomena and their common coverage in undergraduate MCB courses. Owing to their centrality and foundational nature, mitosis and the central dogma (Crick, 1970) are often addressed in K–12 grades, but they are revisited and built upon in more molecular mechanistic detail during undergraduate introductory biology courses. They are then used as foundational principles with which to understand many biological phenomena in upper-division courses such as genetics, cell biology, immunology, developmental biology, and neurobiology. Despite their role as a platform for further learning in biology, the biological mechanisms by which genes are inherited and give rise to traits are often misunderstood by students (Duncan, 2007).

AIM OF THE STUDY

We carried out this study to characterize knowledge integration in undergraduate MCB at different levels of educational development. To explore this, we asked the following research questions:

Research question 1: How do undergraduate students sort, connect, and integrate ideas in the context of learning about transcription, translation, and DNA replication?

Research question 2: What is the nature of connections between ideas in undergraduates’ understanding of transcription, translation, and DNA replication?

METHODS

Study Population

We recruited students for participation in this study on a volunteer basis from one of three courses in the Molecular and Cellular Biology (MCB) program at the University of Arizona: Introductory Biology (N = 22), Molecular Genetics (upper-division, N = 23), and Cell and Developmental Biology (upper-division, N = 8). Most students enrolled in the introductory courses are sophomores or juniors, and most upper-division students are juniors or seniors. We selected these three courses for recruitment, because each course is required for graduation with an MCB major and each course
either explicitly covers the molecular mechanisms of DNA replication, transcription, and translation, and/or makes explicit use of these concepts. Of the 22 introductory students who volunteered for the study, 13 were science majors, two were non-science majors, and seven were MCB majors who are required to continue to the upper-division courses we sampled. Sections of the introductory course are taught by several different instructors with individual design and instructional styles. Because we were interested in exploring the ways in which individual students grapple with learned information, we recruited students from four introductory sections. Each section was taught by a different instructor who routinely used some combination of active-learning techniques such as clicker questions, think–pair–share questions, small-group work, and whole-class discussions. Each upper-division course is taught by a team of instructors, and the teams for both courses intentionally incorporate components of active-learning pedagogy into the courses, including clicker questions, think–pair–share questions, small-group work, data-interpretation studies, analysis of primary literature, and problem-solving sessions.

Discussion of the molecular mechanisms of DNA replication, transcription, and translation in the introductory sections varied slightly based on instructor. These three target molecular mechanisms were reviewed and expanded on in the Molecular Genetics course during units about heredity of information and protein synthesis, structure, and function. The three target mechanisms were then revisited intermittently in the Cell and Developmental Biology course when needed, particularly during units on early embryonic development involving gene expression, the cell cycle and downstream effects of intracellular signaling cascades. By the end of these courses, students are expected to understand the mechanisms of DNA replication, transcription, and translation in significant detail (including, e.g., replication forks, RNA primers, origins of replication, chromosome structure, histones, transcription activators and repressors, directionality of elongation, operons, sites within the ribosome, translation elongation factors, and the process of translation termination) and to apply those mechanisms to cellular and organismal phenomena. In addition to mechanistic detail, all instructors emphasized in some way the rationale behind the use of molecular mechanisms and their connection to overall biological phenomena.

We recruited students for the study through several announcements to the class. We explicitly requested student volunteers who did and did not feel that they understood the course material well to avoid a selection bias as much as possible. The data were collected over the course of four semesters, Fall 2012 to Fall 2014, and a small subset of students participated in interviews in more than one class. Think-aloud interviews were conducted at the end of each respective course in the 2–3-wk period before the final exam. Occasionally, students were interviewed several days after the final exam due to scheduling conflicts. All data collection and analysis was approved by the institutional review board at our university.

**Think-Aloud Interviews**

We conducted semistructured, clinical think-aloud interviews lasting approximately 1 h with each student participant and audio-recorded the interviews. Interviews were conducted by one of six trained interviewers in the research group using a standardized interview protocol (see the Supplemental Material for the protocol). During the interviews, students started by completing a basic concept assessment designed to test conceptual knowledge in a familiar test-like format (see the Supplemental Material). The concept assessment included seven questions using a mix of multiple-choice, short-answer, and fill-in-the-blank questions. Results of the concept assessment (see Figure 1) show a spread of performance, with substantial overlap between the performance of upper-division and introductory students. Concept assessment items were assigned point values, and each students’ concept assessment was graded for correctness. Students’ reported concept assessment scores were calculated as the percentage of points earned out of total possible points. Concept assessment scores were on average higher ($p = 0.006$) in the upper-division group ($89 \pm 1.7\%$, $N = 31$), and variance was lower ($p = 0.015$) than in the introductory group ($76 \pm 3.7\%$, $N = 22$). Owing to the apparent skewness in the distribution of students’ individual percentage concept assessment scores, as revealed by a histogram plot (Figure 1), this tendency toward higher scores in the upper division was confirmed by the Mann-Whitney U-test ($p = 0.006$). (See *Statistical Analysis* section.)

Similarly to Clark and Linn (2013), we used explanations and causal descriptions to probe for “underlying (although often unarticulated) views that shape students’ thinking, explanations, and predictions” (p. 521). During the interview, we asked each student to create explanations and causal descriptions of the processes of DNA replication, transcription, and translation. Specifically, we asked students to describe what they know about the mechanism (or process) of each of the three target mechanisms, including descriptions of what is physically transpiring, the molecular “players,” and where the process is happening within a cell. We did not specify whether students should describe these mechanisms

![Figure 1](image.png)

**Figure 1.** Upper-division students ($N = 31$) perform better than introductory students ($N = 22$) on a basic seven-question concept assessment (two-tailed independent-samples $t$ test, $p = 0.006$).
for eukaryotic or prokaryotic cells; however, most students focused on eukaryotic mechanisms, with only some students choosing to talk about prokaryotic mechanisms as well. Next, we asked students to explain why a cell undergoes that particular process. We asked students to explain their thinking by verbally describing their ideas, and students were often asked to expand their descriptions. Interviewers asked both standardized and individualized follow-up probes to clarify student thinking. We encouraged students to make drawings or sketches when they felt it would aid their verbal description and to be honest about the things they knew and did not know. The interview protocol also included a selected assortment of other tasks, including discussing mutations and problem-solving activities.

One additional task during the interview included the creation of a concept map, for which we asked students to create a free-form concept map using 20 standardized cellular and molecular terms. A brief introduction and instructions on concept mapping were given to students before they began the activity (see the Supplemental Material), and most students were familiar with concept mapping from previous courses. The concept map terms are as follows: cell, phenotype, nucleic acid, gene regulation, inheritance, enzyme, membrane, protein, DNA, transcription, ribosome, replication, mutations, gene, genetic disease, amino acid, neuron, translation, RNA, and skin cell. We asked students to verbalize their thinking as they created a concept map using as many of the given terms as possible. Students were required to label connections between terms with “linker” words, terms, or phrases to describe the relationship. More details about the selection of terms and implementation of the concept maps are provided in the Supplemental Material. After students finished creating their concept maps, we asked them to verbally guide us through the concept maps and describe the relationships they had indicated between terms. Students were also given the opportunity to add terms during this process if they felt it would enhance their concept maps, but most students chose not add terms.

Analysis

Case Study Analysis. We performed case study analysis of the student interviews, using transcripts made from interview audio and any student artifacts (including drawings and notes about gestures) to investigate patterns of reasoning in student interview transcripts. Specifically, we closely examined eight students with diverse levels of conceptual understanding, wrote and discussed claims that might be made about their patterns of reasoning, and compared cases to highlight what seemed to be essential themes and differences between individuals. We will not describe the detailed results from these case studies here, but patterns observed in the case study analysis provided insights to aid in the setup of our coding analysis and pointed us toward portions of the interview protocol that were most useful for addressing our research questions.

Coding Analysis

Mechanism Descriptions. The structure of the interview called for students to first describe the molecular mechanistic events of DNA replication and then describe why a cell would undergo the process of DNA replication. Next, students were asked to describe the molecular mechanistic events of transcription and explain why a cell would undergo transcription. Finally, following this same format, students were asked about translation. We coded each interview transcript for the inclusion of mechanism-appropriate versus mechanism-inappropriate entities. For example, the mechanism of DNA replication would be marked as including “mechanism-appropriate entities” if the student’s description included only entities that are agreed upon by the scientific community as participating in that mechanism, such as DNA polymerase, dNTPs, ligase, or topoisomerase. However, the mechanism was marked as including “mechanism-inappropriate entities” if the student’s description included one or more entities that were inappropriate to the mechanism of DNA replication, such as the involvement of ribosomes or flagella. To investigate the prevalence of misconnections between the three target mechanisms and their roles in the cell, we performed further coding analysis. Student responses to the question “Why does a cell undergo [target mechanism]?” were coded for scientific accuracy or plausibility (scientifically normative connections). Coding analysis was performed independently by three coders, with 78% agreement among coders (number of agreed-upon instances/total number of coded instances). Final codes were the result of consensus among coders. (See the Supplemental Material for full coding scheme.)

Concept Map Connections. Coding analysis of the nature of connections between terms was applied to all student concept maps. A connection was coded as “type 3” if it indicated or implicated a relationship between the terms based on function, mechanism, causation, or an action. A connection was classified as “type 2” if it indicated a relationship between terms based on a vague action, a structural explanation (location, composition, etc.), or a vague role in a process, or if it indicated that one term was a type or example of the other term. Finally, a connection was coded as “type 1” if it was transparently associative (e.g., “associated with”) or was left blank (line or arrow indicating connection but lacking descriptive word or phrase). Terms that were not connected by the student were not coded. Coding analysis was performed independently by three coders with an agreement of 86%. Final codes were the result of consensus among coders. (See the Supplemental Material for coding scheme and an example of a student concept map.)

Iterative Transcript Read-Throughs. We performed systematic read-throughs of all student transcripts for the portions of the interview we identified through case studies as most useful to address our research questions. Read-throughs provided further evidence of themes discovered through case study analysis and were used to triangulate qualitative patterns observed in the coding analysis. All students were given pseudonyms. Quotes from individual students in this paper are identified by pseudonym and course level.

Statistical Analysis. Concept assessment scores were recorded for each student as the percentage of points earned (based on correctness) out of total possible points. Concept map scores were recorded for each student as the average type number (type 1 = 1, type 2 = 2, and type 3 = 3) of the connections drawn by that student. Concept assessment and concept map scores were compared between introductory and upper-division groups using two-tailed,
independent-samples \( t \) tests. Because Levene’s test indicated unequal variances between groups (\( p = 0.015 \) for concept assessment and \( p = 0.003 \) for concept maps), Welch’s unequal variance version of the \( t \) test was used. Histogram plots of individual concept assessment scores and individual concept map scores, by group, suggested nonnormal distributions; therefore, differences between groups were also compared using the nonparametric Mann-Whitney \( U \)-test. Results of the independent-samples \( t \) test were used to describe differences between groups, as there were no instances in which the \( t \) tests were in conflict with the Mann-Whitney \( U \)-tests.

For each of the three target mechanisms (DNA replication, transcription, and translation), student descriptions were coded as including “mechanism-appropriate entities” or “mechanism-inappropriate entities.” Similarly, student connections between the mechanisms and their roles in the cell were coded as “normative” or “non-normative.” Proportions of descriptions including “mechanism-appropriate entities” were compared between introductory and upper-division groups using Fisher’s exact test (which is more appropriate than the chi-square test for the small sample size). Similarly, proportions of “normative” connections between each mechanism and its role in the cell were compared between introductory and upper-division groups using Fisher’s exact test. Statistical significance was declared when the \( p \) value was <0.05; however, actual \( p \) values are reported to provide more detail. Computations were performed using IBM SPSS for Windows, version 22.

RESULTS

Introduction

We have organized our results into two parts. In part 1, we investigate various components of knowledge integration. In part 2, we examine the nature of the connections students make between ideas. Data presented include qualitative coding analysis of interview transcripts (in part 1) and students’ concept maps (in part 2) and samples of student thinking during interviews that exemplify patterns we found in qualitative read-through analysis of transcripts (in parts 1 and 2).

Part 1. Knowledge Integration: Sorting through Ideas about Molecular Mechanisms in Biology

Sorting Molecular Entities into Appropriate Categories. As we previously described, a key component of the theory of knowledge integration is that learners must sort through ideas to create cohesive knowledge networks. One aspect of creating knowledge networks is the restructuring of connections between ideas, including differentiating one idea into two distinct components. Therefore, we first examined the ways in which students sort ideas about molecular entities into mental categories. To explore this, we investigated the ways in which students were discussing molecular entities in their descriptions of the three target mechanisms: DNA replication, transcription, and translation.

We found that students varied in the extent to which they included scientifically normative (mechanism-appropriate) entities when describing molecular mechanisms. For the mechanism of DNA replication, 72% of students were able to create molecular mechanistic descriptions that included only scientifically normative, mechanism-appropriate entities; 69% of students were able to create descriptions of transcription with mechanism-appropriate entities; and 81% of students were able to create descriptions with mechanism-appropriate entities for translation. No statistically significant difference was found between introductory and upper-division students (Fisher’s exact test: DNA replication, \( p = 0.76 \); transcription, \( p = 0.77 \); translation, \( p = 0.27 \)). Figure 2 shows the percentage of students who included only mechanism-appropriate entities in none, one, two, or three of the three target molecular mechanisms (transcription, translation, and DNA replication). More than half of the students interviewed included mechanism-inappropriate entities in one or more of the three target mechanisms. No significant difference was found between introductory and upper-division students (Mann-Whitney \( U \)-test: \( p = 0.53 \)). This was despite the fact that upper-division students had a significantly higher performance on a written concept assessment (see Methods).

When students included a mechanism-inappropriate entity in their descriptions of one of the three mechanisms, 90% were instances in which students included an entity from one of the other two mechanisms (e.g., discussing the role of the ribosome, which is a translation-specific molecular complex, in the process of DNA replication). Coders were careful not to code “mislabeled entities” as inappropriate entities. Examples of this include vague descriptions (e.g., “that thing that comes in and adds the nucleotides”), when students corrected themselves (e.g., “DNA polymerase, I mean RNA polymerase”), or when a student obviously misspoke...
(e.g., switching the words “transcription” and “translation” but describing the intact mechanisms).

These coding results suggest that some students have difficulty sorting molecular entities into appropriate mental categories that reflect the nature of these mechanisms. While the results of this coding analysis showed us whether entities were being appropriately sorted, qualitative analysis of transcripts during read-throughs revealed how the students were using the entities in their mechanistic descriptions and how they were thinking about their mechanisms. In some cases in which students missorted entities, there was evidence that students were conflating mental categories by blurring the boundaries that distinguish the three target mechanisms. This conflation of mental categories resulted in the inappropriate fusion of two or more of the target mechanisms. For example, when one student was asked to describe the molecular mechanism of DNA replication, he responded:

Line 1: “Okay so, from what I know, based off … the cell … so you have the mRNA that left the nucleus and I am sure it gets transcribed by the RNA and as RNA … no as the ribosomes transcribe the RNA … I want to say that it, like, it adds, like, I want to say it adds, like … [trails off into inaudible mumbling].”

Line 2: Interviewer: “What are you thinking about?”

Line 3: “Now I am trying to tie the replication fork in now.”

Line 4: “So … you have … I want to say that the ribosome … like … adds … I want to say, like, temporary base pairs to it?”

Line 5: “I don’t really know what happens there … but all I know is … so, like, you have this when it, like, splits … I know, like, an animal and, like, eukaryotic … they have, like … let’s see … so the DNA … um … so this could be, like, the leading strand and lagging … and then there is something here …”

Line 6: “Something that goes here … I think it’s, um, no RN … cause all I know in the leading strand there is, like … cause, like, I don’t know which enzyme or protein … or whatever it is that … so, like … I think it’s in the lagging strand and something goes along and, like, fills in, like … it adds, like, temporary … I don’t know what it’s called … RNA-tides?” (Jasper, introductory student)

This student’s description indicates that he is aware that the task at hand requires him to create a molecular mechanistic description. He begins his description of the mechanism by using both biological entities and activities. His description includes several entities that are appropriate to the mechanism of DNA replication (replication fork, DNA, leading and lagging strands: lines 3–5). However, he also includes entities and activities like mRNA, ribosomes, mRNA transport out of the nucleus, and transcription, which are all features of the process of protein synthesis (involving the mechanisms of transcription and translation: line 1). The inclusion of both mechanism-appropriate entities and entities belonging to the mechanisms of transcription and translation reveals a conflation, or lack of appropriate differentiation, between the functional modules of DNA replication, transcription, and translation. This phenomenon of conflation between mental categories was observed in several students’ descriptions of the target mechanisms.

In contrast, when students included only mechanism-appropriate entities in their mechanistic descriptions, mental categories appeared to be clearly defined. For example, another student’s description of DNA replication contains no evidence of inappropriate conflation with transcription or translation:

Line 1: “DNA polymerase would come in, but it can’t start by itself so it needs RNA primase to make the primers, to make those three primers in the 5-prime to 3-prime direction, and so when it does that then DNA polymerase just goes in the 5-prime to 3-prime direction, as it is going [the] helicase unwinds the DNA and it separates it.”

Line 2: “And in that direction it just goes fine, just, like, all the way. And then in the 3-prime to 5-prime direction it’s harder because you … it’s like … so RNA primase comes in every so often and adds in primers then DNA polymerase, because it can’t go 3-prime to 5-prime, it kinda goes in the other way and does little chunks at a time” […]

Line 3: “And so it does those in little chunks and those fragments are called Okazaki fragments.” (Bobby, upper-division student)

In this excerpt, the student is, like the previous student, including mechanism-appropriate entities (DNA polymerase, primers, helicase, Okazaki fragments), but unlike the previous student, includes no mechanism-inappropriate entities. Thus, correct sorting of entities seemed to be associated with clear boundaries between mechanisms, as opposed to conflation of mechanisms.

Connecting Biological Ideas. Another key component of the theory of knowledge integration is the creation of connections between ideas, particularly connections that integrate knowledge elements. Therefore, we next explored the connections students made between the mechanical details of a molecular mechanism and that mechanism’s overall function. After students were asked to create a molecular mechanistic description of the three target mechanisms, we asked them to explain why a cell would undergo that particular process.

Analysis of students’ responses revealed a range of connections. Most students were able to create scientifically normative or plausible connections between the target molecular mechanisms and their roles in the cell. Seventy-four percent of students were able to create a scientifically normative connection between the molecular mechanism of DNA replication and its role in the cell; 83 and 88% of students were able to create a normative connection between transcription and translation and their roles in the cell, respectively. No statistically significant difference was found between introductory and upper-division students (Fisher’s exact test: replication, p = 1.0; transcription, p = 0.13; translation, p = 0.39). However, a subset of students created alternative connections between these processes and their roles in cellular function, particularly for DNA replication.
create mechanistic descriptions with entirely mechanism-appropriate entities \((p = 0.50)\). Next, we provide some examples to further illustrate the types of connections students made between mechanisms and their biological role. When Veronica, an introductory student, was asked why a cell would undergo DNA replication, she answered, “Well, it’s going to have to go through mitosis and like split up all of its organelles and everything, so it’s going to need to have a copy of DNA. So it just like makes a replication [copy] of itself, and so it’ll split it in half like two equal [copies].” Here, this student is creating a normative connection between her previously described ideas about the mechanism of DNA replication and the overarching phenomenon of cellular division. Several students created similarly normative connections between the process of translation and protein synthesis, as demonstrated here with another student’s description; Nikki, an upper-division student, says, “To make a functional, to make a protein out of the mRNA so that it [the protein] can function in the cell.”

Although such connections are basic components of these biological phenomena, a significant subset of students described alternative connections (see Table 2). For example, when explaining why a cell undergoes DNA replication, Lydia says:

“Um, to, I guess to make a copy of itself. Or, to do certain duties for example, um, protein to make proteins. Without DNA, you wouldn’t be able to make proteins and without replication you wouldn’t be able to make more DNA in order to make more protein.” (John, introductory student)

Thus, despite the level of mechanistic understanding for DNA replication, this student connected the mechanism to protein synthesis rather than mitosis or meiosis. Coding analysis (see Table 1) showed the number of students with normative or non-normative connections between the mechanism of DNA replication and its role in the cell. Results are broken down by whether the student included entirely mechanism-appropriate entities into his or her description of the mechanism or whether mechanism-inappropriate entities were included. Fisher’s exact test showed no significant association between the ability to make normative connections to the function of the mechanism and the ability to create mechanistic descriptions with entirely mechanism-appropriate entities \((p = 0.50)\).

Next, we provide some examples to further illustrate the types of connections students made between mechanisms and their biological role. When Veronica, an introductory student, was asked why a cell would undergo DNA replication, she answered, “Well, it’s going to have to go through mitosis and like split up all of its organelles and everything, so it’s going to need to have a copy of DNA. So it just like makes a replication [copy] of itself, and so it’ll split it in half like two equal [copies].” Here, this student is creating a normative connection between her previously described ideas about the mechanism of DNA replication and the overarching phenomenon of cellular division. Several students created similarly normative connections between the process of translation and protein synthesis, as demonstrated here with another student’s description of the role of translation; Nikki, an upper-division student, says, “To make a functional, to make a protein out of the mRNA so that it [the protein] can function in the cell.”

Although such connections are basic components of these biological phenomena, a significant subset of students described alternative connections (see Table 2). For example, when explaining why a cell undergoes DNA replication, Lydia says:

“That is a good question ... um, probably to keep the DNA, that is kind of going around in the nucleus, new and have it be able to be used by all these other things that DNA is used for. So there is lots of DNA that can be used for replication, lots of DNA for transcription, lots of DNA to be used for mitosis I guess, too.” (Lydia, introductory student)

In this example, Lydia implies that DNA replication is needed to make sure the cell has “enough” DNA for all the

| Description included mechanism-appropriate entities only | 29 | 9 |
| Description included some mechanism-inappropriate entities | 10 | 5 |

| Description included mechanism-appropriate entities only | 29 | 9 |
| Description included some mechanism-inappropriate entities | 10 | 5 |

*Number of students. \(N = 53\), combined introductory and upper-division students.

| DNA needs to be kept “fresh,” “protected,” “young” | 6/14 | DNA disappears (Rachel, UD) DNA wears out (Brittany, UD) Risky to use original DNA, so make copies (Amy, UD) |
| Need “enough” DNA to make proteins or use for other processes. | 8/14 | Needs lots of DNA for replication, transcription, translation, and mitosis (Lydia, IN) Need extra copies to make proteins (Maria, UD; Amy, UD; John, IN) |
| The cellular “purpose” of the process of DNA replication is to create proteins | 2/14 | Purpose of DNA replication is to make proteins (John, IN; Jasper, IN) |

*Most students displayed more than one heuristic in their non-normative explanations.

*UD, upper-division students; IN, introductory students.
Integrating Ideas to Build a Network of Biological Ideas. Finally, we explored the integration of biological ideas by analyzing the ways in which students were connecting the molecular mechanisms of DNA replication, transcription, and translation to other biological ideas. By analyzing student descriptions of the three target molecular mechanisms, we discovered that a few students immediately analyzed student descriptions of the three target molecular mechanisms of DNA replication, transcription, and translation to other biological ideas. By connecting the molecular mechanisms of DNA replication, transcription, and translation to other biological ideas, students were extending beyond merely incorporating new information into their mechanistic descriptions and instead were making connections they believed to be functionally or causally integrated to their explanations. For example, Sally crafted a description of her understanding of “mutations” when asked to elaborate during the concept-mapping portion of the interview. After connecting mutations to their potential origins during DNA replication, she went on to connect the idea of sequence changes to the role of mutations in creating genetic differences within a population:

“She continues to expand on this idea by connecting it to two relevant biological scenarios in which this principle...”

Here, the student begins by making a connection to chemical signaling, which is a mechanism for cell communication and is the starting condition for many cellular responses like DNA replication. He then creates connections to two related mechanisms, cellular regeneration and wound healing in the context of skin cells. Drawing these two mechanisms into a biologically plausible context (replacing dead skin cells or wound healing), he problematizes the cellular need for DNA replication, and therefore DNA replication. He then creates a connection to regulation and control by explaining the relationship between controlled cellular growth (in wound healing and cellular regeneration) and uncontrolled cellular replication (in the context of cancer). We observe this student extending beyond merely creating a normative connection or two between the molecular mechanism of DNA replication and its role in the cell. Instead, we observe this student’s mental category of DNA replication as being nested within the related biological phenomena of chemical signaling, cellular regeneration and repair, controlled cellular division, and cancer.

Similarly, after describing the molecular mechanism of transcription and the role of RNA as an information messenger for protein synthesis, another student creates a spontaneous connection to regulatory noncoding RNA:

“...but sometimes it does good things, which is why we need skin cells over here, so it will replicate to fulfill the cellular need for DNA replication. And cells have a life span too, and if the cell is old then it needs to be replaced, and cells will signal to each other when we need more cells or whatever.”

By including these additional connections to their explanations, students were extending beyond merely incorporating more information into their mechanistic descriptions and instead were making connections they believed to be functionally or causally integrated to their explanations. For example, Sally crafted a description of her understanding of “mutations” when asked to elaborate during the concept-mapping portion of the interview. After connecting mutations to their potential origins during DNA replication, she went on to connect the idea of sequence changes to the role of mutations in creating genetic differences within a population:

“A mutation is when there is some kind of change in the DNA. And it comes about, I think, usually during replication if there’s an error made by the proteins that are performing the replication. And there’s, like, the proteins that check for errors, there’s... they occasionally mess up. So that’s how mutations come about. They’re just, they change the structure of the protein, sometimes...”

But the replication is a controlled process, it doesn’t just replicate whenever or else it would be cancerous.
holds true; first connecting to malaria resistance in individuals with sickle-cell anemia and then to receptors recognized by HIV:

“In northern Europe there’s, like, this entire deletion of like 32 nucleotides in one gene that codes for a protein that helps the HIV virus recognize cells. So if you get rid of that protein [by way of a mutation], it’s like a receptor protein, by deleting those nucleotides, those people actually become resistant to HIV because HIV can’t bind to the cells. So that’s pretty cool. That’s a good mutation. But usually they do things like they delete proteins, or they change them so that they are not functional, and usually that’s bad.” (Sally, upper-division student)

Through her description, we observe how Sally integrates the ideas of “DNA replication,” “proteins,” “mutations,” and “genetic differences,” and the implications of detrimental versus beneficial mutations that are solidified using two biologically relevant contextual examples. Nuanced connections like those exemplified here by James, Jenna, and Sally suggest that students may begin to build a network of biological ideas at the undergraduate level, integrating their ideas about the target mechanisms with functionally related biological phenomena, principles, and contexts.

**Summary of Part 1 Results.** In summary, we observed that students must develop mental categories with which to appropriately sort the molecular entities involved in each of the target molecular mechanisms. Some students appear to confound one or more of these mental categories, leading to conflated mental models of the molecular mechanisms. Examination of how students connect molecular mechanisms to their overarching roles in the cell suggested that these two aspects of understanding molecular biology are not always linked by students. Finally, we found evidence of some students creating complex knowledge networks that reflect nested and overlapping ideas, suggesting a more nuanced understanding of these topics.

**Part 2. The Nature of Connections**

Through analysis of the previously described results, we uncovered several important differences in the ways in which students created connections. These differences suggested inherent differences in the nature of connections between ideas in MCB. To understand the nature of connections between ideas in undergraduate molecular biology, we took two approaches. First, we examined how students were connecting terms during a concept-mapping task. By analyzing the types of linking words students used to describe the relationship between two concept map terms, we were able to determine differences in types of connections. Second, we qualitatively examined students’ mechanistic explanations of the target mechanisms (DNA replication, transcription, and translation), with the lens of domain-specific reasoning strategies defined in existing literature. Findings included several characteristics of the connections between ideas within the mechanistic descriptions created by the students about DNA replication, transcription, and translation.

**Characterizing the Nature of Connections through Concept Mapping: Associative, Functional, and Causal Connections.** We examined the concept maps created by students during the concept-mapping task of the interviews. This revealed differences in the types of connections drawn between the same ideas (provided terms) by different students. A coding scheme was developed and applied to all student concept maps to systematically look at the nature of connections created between terms (Figure 3, A and B). Connections were coded as type 3 connections when linker words indicated functional, causal, mechanistic, or action-based ideas. Linker words that indicated vague actions or roles in a process, a type or example, or a structural indication (location, composition, etc.) were coded as type 2 connections. And finally, associative or blank connections were coded as type 1.

Results of the coding analysis indicated a significant difference between introductory and upper-division students’ average connection scores (independent-samples t test, p = 0.013). Although roughly half of connection linker terms on introductory students’ concept maps indicated associative connections, only 23% of upper-division concept map connections were classified as associative (Figure 3B). The percentage of type 2 and 3 connections was higher for upper-division students compared with introductory students, 19% higher for type 2 and 8% higher for type 3. It is important to mention that these categories do not reflect the “correctness” of the connections but differences in the reasoning behind why two ideas might be related. It was possible for type 1 and 2 connections to be scientifically normative and for type 3 connections to reflect misconceptions.

After establishing differences in the nature of connections between ideas through the concept-mapping activity, we returned to the students’ descriptions of the molecular mechanisms of DNA replication, transcription, and translation. By looking more deeply into the nature of connections between ideas, we found evidence to support the differences between causal, functional, and mechanistic connections in comparison with associative connections in students’ descriptions of the target mechanisms. Consider the following response when Mandy, an upper-division student, was asked why she included a particular piece of information (the directionality of synthesis of the newly made DNA strand) when describing the mechanism of DNA replication: “Um, I don’t know if I know the answer. I feel like that was just, I don’t know, I’ve just been told that ever since like eighth grade or something, so it’s like, ‘ok, 5-prime, 3-prime.’” After including the same piece of information about directionality in his explanation, Bobby was asked the same question and responded by saying “because on the, uh, 3-prime end there’s a hydroxyl group and that’s where, that’s where the phosphate of the 5-prime end of the next nucleotide, that’s where it attached to. On that 3-prime hydroxyl.” Whereas the first student’s connection between the events of DNA replication and this idea of directionality appears to be based on association, the second student’s same connection appears to be based on functional and mechanistic principles of the molecules involved. Thus, the different forms of reasoning that can be inferred from what students wrote on their concept maps are also evident in their verbal descriptions and explanations.

The results of the coding analysis showed that individual students almost always included connections in more than one category. Furthermore, different students frequently
linked the same two terms with different types of connections. Thus, we hypothesize that the individual knowledge-integration process includes linking of ideas through a variety of reasoning processes, which is likely highly context-dependent and unique to the individual. However, while it may not be essential for all ideas in a knowledge network to be connected by mechanistic reasoning, our data suggest that associative connections are not typically as robust as causal, functional, and mechanistic connections.

Characterizing Connections through Molecular Descriptions: Mechanistic Reasoning. Finally, we provide our findings on the ways that students connected ideas within their descriptions of DNA replication, transcription and translation. Some students created causally connected descriptions, in which entities and activities were used to create a cohesive explanation. A small subset of students not only created productively continuous, cohesive descriptions, but also extended beyond basic causal chaining of ideas by inserting causal additions that reinforced causal connections. However, other students created fragmented explanations that were discontinuous and noncohesive, often including non-normative ideas.

Causal Connections: Mechanistic Reasoning. Machamer, Darden, and Craver propose that mechanisms that explain phenomena are produced by tracing the continuous sequence of events, from setup to termination conditions (Machamer et al., 2000; Russ et al., 2008). Student descriptions that were considered causally connected recounted a sequence of molecular events in which activities are carried out by entities with specific properties. These causally connected descriptions were defined by a continuous sequence of events that often included the temporal and spatial organization of entities and activities, often using activities as a causal driver for the addition of the next piece of the explanation. For example consider this student’s description of the mechanism of DNA replication:

“And so when DNA is to be replicated, it’s double stranded and in order to replicate it the two strands have to be pulled apart from each other forming single stranded DNA. But still connected together. An enzyme called helicase is responsible ... I could have drawn that ... but it, [it] opens up the DNA, it forms the replication fork and then there’s various other proteins [single-stranded binding proteins] that bind to the DNA to prevent it from closing back in on itself, so you want to keep that replication bubble open. I forget what those proteins are called, but they just sit near helicase. I know, just to keep it from closing back up because of the complementary nature of the bases it’s always wanting to close so you need something to prop it open basically.” (James, upper-division student)
This student creates a description of this mechanism by recounting a sequence of molecular events in which activities are carried out by entities with specific properties. Scientifically normative physical entities associated with the mechanism of DNA replication (helicase, replication fork, etc.) are connected through a causal sequence of events. This form of reasoning has been described in the literature as reasoning in which a scientist or a science student uses knowledge about the inherent causal nature of the mechanism to articulate connections between ideas and create a mechanistic description, causally “chaining” through a sequence of ideas (Russ et al., 2008). We observed that some students, like this student, in addition to causally chaining through molecular events, inserted causal additions to build more nuanced explanations (illustrated in the example by underlining). These causal additions appear to reinforce and strengthen causal connections within a chain of molecular events, giving a causal rationale or further explaining the function of a particular connection. It should be noted that explanations including causal additions to a productively continuous causal chain of molecular events included only rare instances of incorporation of mechanism-inappropriate entities. In other words, this more expert form of reasoning did not seem to be associated with mechanism conflation.

Fragmentation: Associative Reasoning. Fragmented descriptions, by contrast, were characterized by frequent disruptions to a chain of molecular events. Consider this excerpt from Jasper’s transcript:

“I don’t really know what happens there … but all I know is … so like, you have this when it, like, splits … I know, like, an animal and, like, eukaryotic … they have, like … let’s see … so the DNA … um … so this could be, like, the leading strand and lagging [drawing] … and then there is something here … something that goes here … I think it’s, um, no RN … ‘cause all I know in the leading strand there is, like … ‘cause, like, I don’t know which enzyme or protein … or whatever it is that … so like … I think it’s in the lagging strand and something goes along and, like, fills in, like, it adds, like, temporary … I don’t know what it’s called … RNA-tides? I don’t know, but I have the image in my head, … but, like …” (Jasper, introductory student)

We describe this explanation as fragmented due to its lack of productive continuity. However, this student does appear to be aware of the task of creating a description of a biological mechanism involves creating a sequenced description of molecular events but is unable to do so. When at a loss for a productive causal connection, this student relies on associative connections to bring in related idea fragments. This connection-building strategy led to frequent disruptions in his chain of events and resulted in a description consisting of weakly associated idea fragments.

Other student transcripts revealed similarly fragmented descriptions of molecular events, in which students used associative connection making strategies. For example, when a different student was asked to describe the same mechanism of DNA replication she said:

“Replication … gosh, it is in the, it’s starts in the nucleus. And then I want to say that it ends in the cytoplasm, and some things that are involved are DNA of course, ribosome, like the larger and the smaller subunit, DNA polymerase, primers, which are RNA primers, helicase, which is another enzyme, and then … so when, it’s when they are separated by … the DNA strand.” (Emma, upper-division student)

Although her description begins with the start and termination conditions of the mechanism, she quickly reverts to listing entities associated with this mechanism. Her list includes both mechanism-appropriate and mechanism-inappropriate entities. As with the first fragmented quote, the majority of this student’s connections appear to be associative in nature. Analysis revealed that these fragmented explanations were closely associated with the inclusion of mechanism-inappropriate entities.

Summary of Part 2 Results. In summary, we observed differences in the ways in which students create connections during the concept-mapping portion of the interview. We further investigated these differences through in-depth analysis of student descriptions of the target molecular mechanisms. Our findings suggest that more robust descriptions were associated with use of mechanistic connections instead of associative connections. In other words, differences in the ways student understand molecular mechanisms can be attributed to differences in the nature of the connections they form between the components of that mechanism. Further, we posit that associative rather than mechanistic connections could serve as a barrier to a deeper understanding of these biological processes.

**DISCUSSION**

In this paper, we provide the first evidence for how the knowledge-integration theory can be applied in undergraduate biology. We have identified key aspects of how students in molecular biology sort, connect, and integrate ideas into a knowledge network. To our knowledge, this study represents the first effort to characterize domain-specific knowledge integration at the undergraduate level. We discovered that students must develop mental categories into which they can appropriately sort the molecular entities involved in the biological processes of DNA replication, transcription, and translation. We characterized the different ways students use associative, heuristic, and mechanistic reasoning when connecting ideas, and we observed that these differences can have consequences for the scientific appropriateness of students’ ideas. We showed how students create nuanced knowledge networks that reflect the multilevel nature of phenomena in molecular biology. Overall, our results support prior work suggesting that knowledge structures are dynamic, emergent, and complex (Brown, 2013; Clark and Linn, 2013). Importantly, our data underscore that knowledge integration is an ongoing process, through which a network of connections are built and refined. Our analysis has revealed that a snapshot of student understanding for a given topic at any point may reveal both scientifically normative and non-normative ideas. Next, we provide our model of knowledge integration, highlighting several of our main findings (see Figure 4).
Model Based on Study Findings

Using the context of the molecular mechanisms of DNA replication, transcription, and translation, our model proposes several features of the process of knowledge integration (i.e., sorting, connecting, and integrating ideas into a knowledge network) in the domain of MCB.

First, we illustrate that fragmentation of ideas decreases as knowledge becomes more integrated. Second, the process of knowledge integration in this domain includes the creation of boundaries between mental categories, depicted as large dotted lines transitioning to solid lines that separate groups of ideas. Third, while overlapping biological ideas may initially result in conflation of mechanisms (“fragmented” section), ideas ultimately must be functionally overlapped in scientifically normative ways for students to create an integrated knowledge network (“nuanced” section). To illustrate our ideas, we describe each of the depicted levels, providing hypothetical examples of student thinking based on our study findings.

In the first “fragmented” section you can see that a hypothetical student, when asked to describe the biological processes of DNA replication, transcription, and translation, names several molecular entities (depicted by the shaded squares). Though mechanistic entities are present, the hypothetical student is unclear which entities belong with which mechanisms. In other words, the student is uncertain where the boundaries lie between the three mechanisms—where one begins and the other ends. For example, while it might be appropriate to include the molecular entity of “DNA” in both DNA replication and transcription, it is not scientifically appropriate to include the molecular entity “ribosome” when describing the mechanics of DNA replication or transcription. We propose that students with “fragmented” descriptions are likely to have undifferentiated mental categories for the three target mechanisms, leading to category conflation or inappropriate idea coalescence. Students with “fragmented” descriptions often struggled to create productively continuous explanations. These student descriptions often have frequent disruptions in their descriptions of the chain of molecular events and often name molecular entities that they associate with the mechanism while being unsure of their specific roles. Without understanding of a chain of molecular events, this student relies on vague and associative connections to attempt to build a mechanistic explanation (depicted by small dotted connections between entities).

In the “transitional” section, we illustrate the beginnings of boundary definition between mental categories for the three target mechanisms for a hypothetical student. At this level, boundaries between mechanisms are still tenuous (dotted lines), and mechanisms may or may not be connected to appropriate biological contexts (labeled “transcription,” tentative label for “DNA replication,” and missing label for “translation”). Because boundaries between mechanisms are still ill-defined, occasional errors in the inclusion of mechanism-inappropriate entities persist. For example, this hypothetical student has discussed the role of RNA polymerase both in DNA replication and transcription, when DNA polymerase is the mechanism-appropriate entity...
for the activity of DNA polymerization. While vague and associative connections are still present, “transitional” descriptions include causal or functional connections between molecular events, depicted as solid arrows.

In the third section, student descriptions can be described generally as “connected.” A hypothetical student has appropriately sorted all entities into the appropriate categories for DNA replication, transcription, and translation, and these ideas have been connected to (or nested within) scientifically normative roles in the cell (protein synthesis and cellular division). Molecular entities are connected by corresponding activities that are temporally and spatially defined within the mechanistic chain of molecular events (solid arrows).

And finally, in the fourth section, we illustrate a hypothetical student’s “nuanced” network of biological ideas. Here, the student has included mechanism-appropriate molecular entities into her or his description of the three target mechanisms, defined appropriate yet flexible boundaries between mental categories, and described a productively continuous chain of molecular events in which entities have corresponding temporal and spatial activities. In addition, this hypothetical student adds functional or causal connections to relevant biological phenomena and integrates several ideas to describe a nuanced and overarching biological principle.

Our data suggest that conflation of molecular mechanisms is not uncommon among undergraduate students. This is perhaps not surprising, given the complex, multilevel nature of MCB. In particular, our model points to the importance of functional modules as a level of organization. Although functional modules have been previously described as inherent features of biological systems (Hartwell et al., 1999; Hofmann et al., 2006), we propose that they also serve as mental organizers. Students must integrate knowledge about the mechanics of a molecular mechanism with its overarching role in the cell or organism, thereby housing the features and mechanics of a molecular mechanism within an overarching biological function. We hypothesize that experts within the field of MCB use functional modules as mental organizers, which works to drive the appropriate categorization of new knowledge into appropriate mental categories. As our model depicts, we suggest that, for knowledge to become more integrated, students must use functional and causal drivers as organizing pillars to create an integrated network of biological ideas. Overall, the process of mental category development appears to be a complex and sensitive process. This category development requires that category boundaries be well enough defined to differentiate ideas but flexible enough to allow dynamic relationships between these ideas in order to construct a network of biological ideas.

Organizing Ideas in Molecular and Cellular Biology

Within the field of molecular biology, a biological mechanism and its role in the cell are inseparably linked. Our results suggest that, in an educational setting, these connections are not always apparent or intuitive to students. Whereas experts in the field of MCB might find it difficult to separate molecular mechanisms from their functions, our data suggest that the “mechanism” and “function” aspects of functional modules can be separated for students learning about molecular mechanisms. Our results suggest that students’ mental categories for molecular mechanisms do not necessarily develop in parallel to the development of their understanding of the functions of the mechanisms. Although many students could describe molecular events mechanistically, in some cases, these descriptions seemed to be disconnected, that is, mechanisms without a cause.

A notable number of students struggled to connect the mechanism of DNA replication to its role in the cell. On the basis of prior work suggesting that it can be difficult for students to make connections between physical and ontological levels (Marbach-Ad and Stavy, 2000; Duncan, 2007; Duncan and Tseng, 2011), and more specifically between the molecular and cellular levels (Van Míl et al., 2013), we suggest that it may be difficult to build connections between the molecular level (mechanism of DNA replication) and cellular biophysical level (process of mitosis and meiosis). By contrast, the molecular mechanisms of transcription and translation may be more easily connected to gene regulation and protein synthesis, which are also at the molecular level.

While our results indicated disconnects like these in some students’ understanding, we also explored the more nuanced connections that students made. The “core features” previously described in the literature as underlying organizers for expert knowledge structure (Chi et al., 1981; Chi and Koeske, 1983; National Research Council [NRC], 2000) parallel the abstract, function-based principles we observed being used by a few students to nest and overlap concepts to create complex causal explanations. The mental categories that we describe bear similarity to what the expertise literature has defined as “chunks.” These “chunks” are collections of ideas organized based on underlying features and allow for rapid retrieval of information and enhanced problem-solving abilities (De Groot, 1965; Chase and Simon, 1973; NRC, 2000). Experts must organize ideas about molecular mechanisms by underlying functional features, “chunking” ideas and creating complex networks between chunks. Thus, students of molecular biology must build a functionally connected knowledge network that includes organization into functional modules. Further, though it has been hypothesized that having an integrated knowledge network may contribute to the ability that experts have to solve novel problems (Linn, 1995; Jonassen, 2000), this idea has not been fully explored empirically. So, while we theorize that students in our study with a more nuanced network of ideas might be better prepared to approach scientific problems, it is not within the scope of this study to claim that integrated knowledge will improve transfer.

Developing Integrated Knowledge

Our results suggest that learning is occurring between introductory and upper-division courses, as demonstrated by a significant difference in performance on a basic concept assessment. Our results also demonstrate a difference in the type of connections, on average, between introductory and upper-division students. These results suggest that more novice students in introductory courses may be relying on associative connections between ideas and that, by upper-division courses, students are beginning to move toward more functional, causal, and mechanistic connections by reconnecting or modifying existing connections. Further, the explanations provided by students in our study, at both introductory and upper-division levels, primarily included
mechanistic reasoning. Though heuristic (rule-based) explanations were occasionally used, we saw little evidence of teleological explanations. Overall, our results suggest that explanations in our study population were quite different from younger students in prior studies (Southerland et al., 2001; Duncan and Reiser, 2007; Bolger et al., 2012).

However, some aspects of forming an integrated knowledge network remain challenging for students across courses. This was demonstrated in our results showing no significant differences between introductory and upper-division student populations’ ability to sort entities into mechanism-appropriate mental categories and connect molecular mechanisms to their roles in the cell. The fact that conflation of categories persists among the population of upper-division students suggests that the creation and modification of mental categories, a process by which knowledge about the central dogma is organized, is continuous and ongoing as education proceeds. Although our data cannot suggest patterns in knowledge integration over time or make hypotheses about potential learning progressions, we do suggest that longitudinal studies are required to track how sorting, connecting, and integration of ideas change over time for individual students. Further, while it might be attractive to think of our proposed model as steps along a learning progression, our data do not suggest this as a linear set of steps that students must go through to gain expertise.

We were able to track a small number of students from their introductory courses to upper-division courses, because a few students volunteered for interviews in both their introductory and upper-division courses. Examination of transcripts from these students suggests, as you might expect, that some students showed significant improvement in the way they connected ideas between interviews conducted in the introductory course and interviews conducted in the upper-division course. However, we also saw evidence to suggest that some students, while showing some improvements in certain areas, also developed significant misconnections between introductory and upper-division courses that led to conflation of DNA replication, transcription, and translation. These examples hint at the nonlinear nature of learning in which many students may take overarching, productive steps in building scientifically normative knowledge networks, but the day-to-day restructuring and reorganizing of ideas is in fact a complex and dynamic process. Future studies will explore the impact of these ideas on conceptual change over time.

**Instructional Implications**

Our results suggest that some students have difficulties with appropriately differentiating mental categories for the biological mechanisms of DNA replication, transcription, and translation. Traditionally, in introductory courses and some upper-division courses, these three mechanisms have been taught within a narrow window of time, often an intentional design with the instructional objective of comparing and contrasting these similar processes. However, we suggest that instructional strategies that embed these three mechanisms within “big picture” contexts for each mechanism could be more effective in helping students to develop functional modules as mental organizers. By first creating a “need to know,” students may be more likely to make causal connections between mitosis and DNA replication, gene regulation and transcription, protein synthesis and the details of the mechanism of translation.

We do not consider this persistence of conceptual difficulties into upper-division courses to be surprising, as similar results have been found in other studies that compare introductory and upper-division students (Songer and Mintzes, 1994). In fact, we suggest that it provides further evidence for the previously established idea that knowledge integration takes time (Clark and Linn, 2003). Introductory courses in our study spent between 1 and 2 wk of instructional time on DNA replication, transcription, and translation; there is a brief review of these concepts in upper-division courses before students are asked to connect these ideas to related topics. Therefore, it is not surprising that incomplete differentiation of mental categories or non-normative connections are not uncommon, even among upper-division MCB majors. Instead, we view our results as additional evidence suggesting that coverage of fewer topics in undergraduate biology courses is needed if students are to develop deeper understanding.

Our findings also suggest that instructors should be careful when designing assessments on these topics. Assessments should reach into the types of connections students are making between ideas and the ways in which their ideas are organized. By testing underlying reasoning patterns, instead of learned vocabulary, we can communicate to students the importance of knowledge organization and integration in addition to the ability to recite declarative information. To be productive, assessments must move away from tasks that allow students to rearrange causal language associated with surface features of mechanisms or to create answers that sound like expertise without revealing whether underlying knowledge remains organized by associative reasoning.

The process of knowledge integration is neither linear nor straightforward. As instructors, when we observe non-normative ideas in our classrooms, it is only natural to return to the original lesson plan with a critical eye. However, our results support the idea that student knowledge integration is an essential, internal, and individual process by which students need to actively reorganize and reinterpret acquired knowledge. As a result, when assessments reveal non-normative ideas, we must be cognizant of what the particular non-normative ideas might reveal about underlying connections and organization in the students’ knowledge structures. We must be aware that these non-normative connections often are not simply due to a lack of detail or effort on the part of the student but may reflect active restructuring of knowledge that sometimes creates inappropriate connections. These misconnections may result in ideas that are commonly referred to as “misconceptions.” However, the reasoning patterns that result in these misconceptions may actually represent a misapplication of reasoning that could lead to a productive connection in another context. We urge instructors to cultivate intentional awareness of how students are sorting ideas, creating connections, and integrating knowledge, as these are integral to the dynamic and continual process of learning.

**ACKNOWLEDGMENTS**

We acknowledge Dr. Phillip Chapman, Professor Emeritus, Colorado State University, for his guidance in our statistical analysis.
REFERENCES


Scientist Spotlight Homework Assignments Shift Students’ Stereotypes of Scientists and Enhance Science Identity in a Diverse Introductory Science Class

Jeffrey N. Schinske,* Heather Perkins,* Amanda Snyder,* and Mary Wyer†
†Biology Department, De Anza College, Cupertino, CA 95014; ‡Psychology Department, North Carolina State University, Raleigh, NC 27695

ABSTRACT
Research into science identity, stereotype threat, and possible selves suggests a lack of diverse representations of scientists could impede traditionally underserved students from persisting and succeeding in science. We evaluated a series of metacognitive homework assignments (“Scientist Spotlights”) that featured counterstereotypical examples of scientists in an introductory biology class at a diverse community college. Scientist Spotlights additionally served as tools for content coverage, as scientists were selected to match topics covered each week. We analyzed beginning- and end-of-course essays completed by students during each of five courses with Scientist Spotlights and two courses with equivalent homework assignments that lacked connections to the stories of diverse scientists. Students completing Scientist Spotlights shifted toward counterstereotypical descriptions of scientists and conveyed an enhanced ability to personally relate to scientists following the intervention. Longitudinal data suggested these shifts were maintained 6 months after the completion of the course. Analyses further uncovered correlations between these shifts, interest in science, and course grades. As Scientist Spotlights require very little class time and complement existing curricula, they represent a promising tool for enhancing science identity, shifting stereotypes, and connecting content to issues of equity and diversity in a broad range of STEM classrooms.

INTRODUCTION
Whether or not we consciously register the impacts of this messaging, we are regularly bombarded with information regarding the types of people who work in science, technology, engineering, and mathematics (STEM). From television shows and movies to websites, news articles, and advertisements, the media recurrently conveys images of who does science, more often than not showcasing a relatively narrow view of science and scientists. Setting the media aside, perhaps we need look no further than our own classrooms to understand the ways scientists are portrayed. Many students are likely to get their earliest and most direct experiences with “real” scientists when attending college STEM classes—classes taught by a mostly white, mostly male faculty nationwide (National Science Foundation, 2013). Our textbooks, in the very rare instances they connect content to discussions of specific scientists, can tend to focus the most attention on individuals matching common scientist stereotypes (e.g., Darwin and Mendel in Reece et al., 2014). Even our classrooms themselves may, through their physical layouts and decorations, convey messages regarding who can participate in STEM (Cheryan et al., 2009). We might wonder, then, what are the impacts of these recurrent messages on students enrolled in postsecondary STEM classes, particularly in the increasingly diverse classroom environments of the United States? And what, if anything, might faculty do in response to this messaging?

Pat Marsteller, Monitoring Editor
Submitted January 15, 2016; Revised June 11, 2016; Accepted June 14, 2016
DOI:10.1187/cbe.16-01-0002
The online version of this article contains supplemental material at www.lifesci.org.
*Address correspondence to: Jeffrey N. Schinske (schinskejeff@fhda.edu).
© 2016 J. N. Schinske et al. CBE—Life Sciences Education © 2016 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.
Scientist Stereotypes Impact Persistence and Success in STEM by Influencing Science Identity, Sense of Belonging, and Stereotype Threat

The messages we convey to students, either intentionally or unintentionally, regarding who does science can influence students’ stereotypes of scientists. Many lines of evidence point to the importance of these stereotypes in shaping students’ sense of belonging in STEM, with implications for persistence and success in STEM programs. For example, stereotypical representations of scientists in the media (Tanner, 2009; Cheryan et al., 2013; DeWitt et al., 2013; Martin, 2015) and in classroom decorations (Cheryan et al., 2009) have the potential to reduce interest in STEM fields among women and people of color. On the other hand, a variety of studies suggest students are more likely to pursue majors and careers in STEM if they agree with certain “positive” stereotypes of scientists (Beardslee and O’Dowd, 1961; Wyer, 2003; Schneider, 2010). Our own work further suggests that holding counterstereotypical images of scientists might be an important factor in predicting success in science classes (Schinske et al., 2015).

These findings illustrate the importance of science identity, a sense of belonging, and stereotype threat in determining persistence and success in STEM classes. Identity refers to the extent to which we view ourselves as a particular “kind of person” (Gee, 2000), with science identity more specifically referring to whether we see ourselves as scientists. If students hold stereotypes that portray scientists as a different “kind of person” than themselves, those students might conclude they are not “science people.” This mismatch between a student’s personal sense of identity and a science identity can hamper persistence in STEM (Seymour and Hewitt, 1997; Brickhouse et al., 2000). Harborings views of scientists that differ from students’ perceptions of themselves could also cause students to feel as though they do not belong in science. The extent to which students feel a sense of belonging similarly correlates with levels of achievement and motivation in school settings (Goodenow, 1993; Roeser et al., 1996).

Feeling that one differs from stereotypical descriptions of people in a particular field of study can additionally hinder achievement in that field due to stereotype threat. Under stereotype threat, students harbor an often subconscious fear of confirming a negative stereotype about their groups (Steele, 1997). For example, students of color, women, and first-generation college students might fear confirming a stereotype that their groups are not good at science due to a perception that scientists are white men from privileged, highly educated backgrounds. This threat can undermine engagement and performance, even among students who are otherwise well qualified academically (Steele, 1997). Even subtle cues involving a lack of women or people of color visually represented in an academic environment or on a flyer can trigger dramatic reductions in interest and performance due to stereotype threat (Inzlicht and Ben-Zeev, 2000; Purdie-Vaughns et al., 2008). More specific to science contexts, stereotype threat has been described as a significant factor in predicting interest, persistence, and success in STEM majors, especially for women and students of color (Hill et al., 2010, chap. 3; Beasley and Fischer, 2012). Interventions that remove the conditions that trigger stereotype threat can reduce or even entirely eliminate achievement gaps between women and men or between students of color and white students in test scores and course grades (e.g., Steele and Aronson, 1995; Good et al., 2003; Cohen et al., 2006).

What Can Faculty Do in STEM Classes to Broaden the Image of the Scientist?

Given the evidence suggesting that stereotypes of scientists impact persistence and success in STEM, efforts to feature counterstereotypical images of scientists have the potential to narrow equity gaps and broaden participation in STEM. Stereotypes of scientists are malleable (Cheryan et al., 2015), and previous work suggests that providing counterstereotypical messaging could enhance interest and success in STEM among underserved populations of students (McIntyre et al., 2004; Steinke et al., 2009; Cheryan et al., 2013).

One common strategy for introducing counterstereotypical images of scientists to students is to increase the prevalence and visibility of diverse STEM “role models”—individuals who students may choose to emulate. Marx and Roman (2002) describe how role models are chosen through “selective, social comparison whereby certain attributes are copied and others are excluded.” Because comparisons of social similarity may involve the visible personal characteristics of potential role models, many studies have focused on the potential benefits of gender or race/ethnic-matched role models. For example, the presence of female role models has served to mitigate stereotype threat and boost math performance among female students (Marx and Roman, 2002; Marx and Ko, 2012). In terms of race/ethnicity, both white and nonwhite students tend to select race/ethnic-matched career role models (Karunanyake and Nauta, 2004), and having a race/ethnic-matched instructor role model has been shown to correlate with student success (Dee, 2004; Fairlie et al., 2011).

While these results would suggest placing a priority on seeking out gender/race/ethnic-matched role models for STEM students, other studies have failed to find distinct benefits of role models who match students’ own races/ethnicities and genders (Ehrenberg et al., 1995; Maylor, 2009; Phelan, 2010). Perhaps explaining these discrepancies, Marx and Roman (2002) point out that the attributes important to seek in a role model will ultimately be those attributes of importance to the individual choosing the role model (e.g., the attributes considered important by students). Because social identities are informed by many different factors, and individuals have multiple identities that resonate in different contexts (Gee, 2000), it might be difficult to predict which role model attributes will be most important in encouraging students to form a science identity. Buck et al. (2008) provide guidance in this area in finding that students needed to identify someone “who cared about them and shared common interest/experiences” in order for role models to be effective. This work implies that faculty interested in enhancing students’ science identity and sense of belonging in STEM should, in addition to identifying diverse role models in terms of gender/race/ethnicity, place a priority on featuring individuals to whom students might personally relate, based on interests and experiences.

Moving from Identifying Role Models to Showcasing Possible Selves

The concept of “possible selves” might represent a more useful and precise way to think of counterstereotypical examples than
does the concept of “role modeling.” Possible selves refer to everything that each of us “is tempted to call by the name of me” (James, 2005) or the set of “individually significant hopes, fears, and fantasies” that define oneself (Markus and Nurius, 1986). Individuals can reflect upon their own possible selves, and these possible selves are understood to influence motivation and future behavior (Markus and Nurius, 1986). Students weigh their possible selves in constructing school identities, and these interactions between possible selves and academic identities mediate the potency of stereotype threat (Steele, 1997; Oyserman et al., 2006). Possible selves more specifically play an important role in the development of a science identity (Hunter, 2010), and students’ “possible science selves” might help explain career choices in STEM (Steinke et al., 2009; Mills, 2014). Taken together, this implies students’ science identities and resistance to stereotype threat might be enhanced if they see their own possible selves reflected in STEM. This highlights a subtle but important difference between the concepts of role models and possible selves. Compared with featuring scientist role models that represent people students are expected to become more like, seeing one’s possible self in a scientist would involve seeing someone in science you already are like.

Goals and Scope of This Study
Given the evidence that counterstereotypical perceptions of scientists are important in diverse science classrooms (Schinske et al., 2015) and that viewing one’s possible selves in science might enhance science identity (Hunter, 2010; Mills, 2014) and mitigate stereotype threat (Oyserman et al., 2006), we developed and evaluated a classroom intervention to introduce students to counterstereotypical examples of scientists. In evaluating the intervention, which we call “Scientist Spotlights” (see Methods), we sought to explore the following four hypotheses.

Hypothesis 1: Scientist Spotlights will shift students’ descriptions of scientists toward nonstereotypical descriptions.
Hypothesis 2: Scientist Spotlights will enhance students’ ability to see their possible selves in science by enhancing their ability to relate to scientists.
Hypothesis 3: Shifts in scientist stereotypes and relatability of scientists will correlate with students’ interest in science.
Hypothesis 4: Shifts in scientist stereotypes and relatability of scientists will correlate with course grades.

METHODS
Below we review the development of the Scientist Spotlight intervention, the study context, and our mixed-methods analysis of student essays and quantitative surveys to evaluate the intervention.

Development of Scientist Spotlights in a Diverse Community College Biology Classroom
We developed Scientist Spotlights as regular, out-of-class assignments both to introduce counterstereotypical examples of scientists and to assist in the coverage of course content while requiring little class/grading time. Featured scientists were selected to 1) present diverse perspectives on who scientists are and how science is done and 2) match the content areas being covered at the time of each assignment. In each Scientist Spotlight, students reviewed a resource regarding the scientist’s research (e.g., a journal article or popular science article) and a resource regarding the scientist’s personal history (e.g., an interview, Story Collider podcast, or TED Talk). Because these assignments included the review of materials that introduced course content to students, they replaced weekly textbook readings. One of the Scientist Spotlights assigned to students read as follows:

Ben Barres is a Stanford professor of neurobiology. He studies diseases related to signaling in the nervous system, and in particular the roles of supporting cells around neurons. Dr. Barres is also a leader in science equity and the effort to address gender gaps. He is uniquely positioned to address these issues, since he has presented both as a female and a male scientist at different times in his career.

1. View the Wall Street Journal article about Ben Barres by clicking here (Begley, 2006).
2. Then, review Dr. Barres’ article in the journal Nature by clicking here (Allen and Barres, 2009)

(If you are interested in hearing more from Ben Barres, you can search for him on YouTube. He has some videos on his research and also on his experiences as a transgender person.)

After reviewing these resources, write a 350 word or more reflection with your responses to what you saw. You might wish to discuss:

1. What was most interesting or most confusing about the articles you read about Dr. Barres?
2. What can you learn about neuron signaling (action potentials, synapses, supporting cells) from these articles?
3. What do these articles tell you about the types of people that do science?
4. What new questions do you have after reviewing these articles?

The above example was assigned before a unit on neuron signaling and therefore assisted in the introduction of content in that area. The writing prompts were aimed at creating opportunities for metacognition (Tanner, 2012). Prompts changed slightly from one assignment to the next, but the third prompt about the “types of people that do science” was always included. A photograph of the featured scientist was also included with each assignment. Students submitted responses to Scientist Spotlights through an online course-management system (Moodle), and submissions were scored only for timeliness and word count.

Study Design
We used a quasi-experimental, nonequivalent-groups design (Shadish et al., 2002; Trochim, 2006) to evaluate Scientist Spotlights in a Human Biology course at a diverse community college during the Fall 2013–Fall 2015 academic terms. Human Biology is a one-quarter lecture/lab general education course open to any student, but targeting transfer students and those with interests in human health careers. Students in five sections of Human Biology during that time period completed Scientist Spotlights on a weekly basis (hereafter “Scientist Spotlight
Homework” students). Each Scientist Spotlight was worth 10 points, so the assignments (n = 10) contributed a total of 100 points to the final course grade (865 points in the whole course). Efforts were made to attend to multiple axes of diversity when selecting scientists to feature, with special attention to the racial/ethnic diversity of students in these classes. Half of the weeks featured female scientists and seven out of 10 weeks featured at least one nonwhite scientist. Occasionally, more than one scientist was featured during a Scientist Spotlight assignment. Selected scientists represented diverse socioeconomic backgrounds, gender identities, interests outside science, paths to careers in science, temperaments, ages, sexual orientations, and countries of origin. Supplemental Material, part A, lists the names of individuals featured in Scientist Spotlights during this study. The full set of 10 Scientist Spotlight assignments, including readings and resources, is available by request to the corresponding author.

During the same time period, students in two sections of Human Biology did not perform Scientist Spotlights. Instead, those students completed comparable metacognitive online assignments (example in Supplemental Material, part B) based on popular science articles and journal articles compiled in a course reader (hereafter “Course Reader Homework”) students. Although no explicit instruction regarding scientist stereotypes was mentioned during class while highlighting certain research findings related to course content. Students saw photographs of all three scientists and watched brief videos featuring Dr. Cobb and Dr. Shubin but did not perform any individual/group work or metacognitive activities surrounding these scientists.

Quasi-experimental approaches, by definition, lack randomization in assigning participants to groups (Shadish et al., 2002; Trochim, 2006). As such, students self-selected into Human Biology course sections and the instructor (J.N.S.) selected sections in which to implement Scientist Spotlight versus Course Reader Homework. While nonrandom assignment to groups can limit researchers’ ability to infer causal connections between interventions and outcomes, quasi-experimental approaches can still provide robust and valuable insights and offer advantages over randomized experiments in certain contexts (Shadish et al., 2002). We attempted to ensure as much equivalence as possible between groups in that all classes adhered to the same curricular expectations, were taught at similar times of the day in similarly arranged classrooms, and used the same types of in-class activities. The same faculty member (J.N.S.) served as instructor for all of the course sections involved in this study, though one Course Reader Homework section was cotaught by another faculty member. We controlled for various student-level differences between groups during statistical analyses and used these “weighted means” in evaluating our hypotheses (see Methods and Supplemental Material, part E). It should be noted that, in the analyses that follow, we consider students as the experimental units. This was considered most appropriate in this instance, because Scientist Spotlights were designed to interact with individual students in different ways, raising interest in students as individual observations. We do, however, control for course section in analyses to account for trends based on grouping at the class level.

Student Population
This work was conducted at a large (~22,000 students) California community college that is a designated Asian American and Native American Pacific Islander–Serving Institution (AANAPISI). The majority (59%) of students come from low-socioeconomic status (low-SES) families and the majority (66.2%) indicate the educational goal of transferring to a 4-year institution. Approximately 20% of Human Biology students state the intention of majoring in biology. Forty-six percent of students report that Human Biology is the first college science class they have taken, and 13% of students report that Human Biology is the first science class they have ever taken at any level.

A total of 364 students initially enrolled in the five sections of Human Biology that completed Scientist Spotlight Homework (n = 73 students per class). One hundred thirty-nine students initially enrolled in the Course Reader Homework sections (n = 70 students per class). However, 26 students from Scientist Spotlight Homework classes and 13 students from Course Reader Homework classes dropped the course within the first 2 weeks of class, leaving 338 students as the final enrollment for Scientist Spotlight Homework sections and 126 students in Course Reader Homework sections.

The table in the Supplemental Material, part C, compares the demographic characteristics of students in these classes. We defined “underserved” racial/ethnic groups as those groups that have persistently entered STEM majors at lower rates compared with their prevalence on campus and experienced comparatively lower success rates in STEM classes. This included students identifying as Latino/a, Black, Native American, Filipino/a, Pacific Islander, and Southeast Asian (e.g., Vietnamese, Laotian, Cambodian, Indonesian). The majority of Scientist Spotlight and Course Reader Homework students identify as members of underserved groups (Supplemental Material, part C). Students in these Human Biology classes identified 25 different first languages spoken, with English, Spanish, and Vietnamese representing the most common first languages spoken.

Assessment of Scientist Stereotypes and Possible Science Selves through Short-Essay Surveys
In evaluating Scientist Spotlights, we used a mixed-methods approach in which we reviewed short-essay responses from students for context and themes and then coded student responses into categories for quantitative analysis. Two essay prompts were used. The first prompt was designed to address hypothesis 1 by eliciting students’ stereotypes of scientists. This prompt read, “Based on what you know now, describe the types of people that do science. If possible, refer to specific scientists and what they tell you about the types of people that do science” (hereafter “stereotypes prompt”). This prompt was described and its validity was explored by Schinske et al. (2015). The second prompt was developed as an exploratory method for assessing students’ possible selves in science. That is, assessing whether students perceived scientists as reflecting their possible selves, and if so, what aspects of themselves they saw reflected in scientists (hypothesis 2). We chose to approach this topic by surveying the extent to which students could
“personally relate” to scientists. The prompt consisted of the challenge statement: “I know of one or more important scientist to whom I can personally relate,” followed by a Likert scale including “agree,” “somewhat agree,” “somewhat disagree,” “disagree,” and “I don’t know.” Following the Likert scale, students were instructed: “Please explain your opinion of the statement” (hereafter “relatability prompt”). This prompt was developed and face validity was established through multiple quarters of testing in class and informal talk-aloud trials with students. Even though an “I don’t know” response was essentially the same as “disagree” when students responded whether they knew of one or more relatable scientists (see also Results), we found it important to include an “I don’t know” option. Some students were more comfortable circling “I don’t know” than “disagree,” which sounded like a “wrong” answer to them.

These two prompts were printed on one side of a sheet of paper, so students had approximately half a sheet to respond to each prompt. J.N.S. provided the surveys to students on the first and last days of each Human Biology course, telling students, “I am very interested in students’ ideas about science and scientists, so I appreciate you taking 5–10 min to respond to these prompts. There are absolutely no right or wrong answers and there’s nothing I would like more than to see many different thoughts on the topic. Your responses will not be graded and will not be reviewed in connection with your name.” Though responses were not graded, students received five points (out of 865 course points) for participating and completing surveys. When looking for shifts in attitudes about scientists in these surveys, only papers from students who submitted both beginning- and end-of-course responses were considered. As preliminary results suggested students in Scientist Spotlight Homework classes were adopting new attitudes regarding scientist stereotypes and the relatability of scientists, we were interested in whether those shifts would be maintained over time. To assess these shifts longitudinally, J.N.S. sent an online survey that included the stereotypes and relatability prompts to Scientist Spotlight Homework students approximately 6 months after the end of class.

Analysis of Students’ Descriptions of Scientists
We anonymized and randomized student papers and followed the procedures of Schinske et al. (2015) to categorize responses to the stereotypes prompt. While reviewing student responses, we recorded the words, phrases, and names students used to describe scientists, and tallied the frequencies of those descriptions among the papers. Exemplar quotes were selected to represent the most common themes and provide context. Pseudonyms were used in place of student names to protect anonymity. Students’ descriptions of scientists were then coded as Stereotypes, Nonstereotypes, or Fields of Science. Following our previous work (Schinske et al., 2015), we defined Stereotypes as any widely represented descriptions of scientists matching stereotypes uncovered by Mead and Metraux (1957). Nonstereotypes included less commonly used descriptions of scientists not reported in that previous work. Fields of Science included names of science fields or career types (e.g., biologist). We previously demonstrated that independent reviewers reliably code descriptions as Stereotypes (0.86 interrater correlation) and Nonstereotypes (0.89 interrater correlation; Schinske et al., 2015). We recorded the number of descriptions from each category for each student, then converted those numbers into percentages out of total comments (e.g., percent of Stereotypes out of all comments) to partly control for differences in the lengths of responses between students.

Changes in the proportions of Stereotypes and Nonstereotypes were analyzed using repeated-measure analysis of covariance (RM-ANCOVA). Proportions of Stereotypes/Nonstereotypes acted as dependent variables, with time (beginning vs. end of course) and treatment (Scientist Spotlight Homework vs. Course Reader Homework) input as between-subjects factors. Gender, race/ethnicity (categorized as traditionally underserved vs. traditionally well served), and course section were used as covariates.

Analysis of Students’ Ability to Personally Relate to Scientists
We reviewed short-essay responses to the relatability prompt and transcribed each of students’ statements (e.g., “Don’t know any scientists,” “Relate to musician scientist,” “Relate to Rosalind Franklin”) into the top of a spreadsheet. As those statements reappeared in subsequent papers, we tallied the appearance of the statements in the spreadsheet. Exemplar quotes were selected to represent the most common themes and provide context for why students could or could not personally relate to scientists.

Changes in students’ relatability Likert-scale selections from the beginning to the end of the course, were analyzed using RM-ANCOVAs. Relatability Likert scores acted as the dependent variables, with time and treatment input as between-subjects factors. Gender, race/ethnicity, and course section were used as covariates.

Analysis of Student Interest in Science and Collection of Demographic Information
The exploration of hypothesis 3 required comparing shifts in students’ stereotypes of scientists and ability to relate to scientists to shifts in science interest. To monitor student interest, during the first and the last weeks of class, students completed an online survey (Supplemental Material, part D). The survey included eight quantitative items adapted from the Student Assessment of their Learning Gains Survey (Seymour et al., 2000), which were reshaped into the “Science Interest” scale. Students responded to prompts such as “Presently I am enthusiastic about this subject” on a five-point Likert scale, ranging from “not at all” to “a great deal.” Supplemental Material, parts G and H, provide details regarding how the Science Interest scale was derived from these items. In separate questions, students indicated whether they were majoring in biology or another STEM field and whether they had taken previous science classes (Supplemental Material, part D). As we also wished to look for interactions involving student demographics, the final page of the surveys asked students to identify their gender and racial/ethnic identities and first spoken language. Students received five participation points (out of 865 course points) for completing these quantitative surveys.

Prior work suggested broader student outcomes, like grades and interest in science, relate to holding nonstereotypical views of scientists (Schinske et al., 2015) and developing possible science selves (Steinke et al., 2009; Mills, 2014). We therefore created categorical variables to distinguish students
who exhibited these characteristics. Specifically, we compared end-of-course with beginning-of-course values to categorize students as either decreasing versus not decreasing in their proportion of Stereotypes, increasing versus not increasing in their proportion of Nonstereotypes, and increasing versus not increasing in relatability. The relationships between each of these categorical variables and Science Interest were tested in a $2 \times 2 \times 2$ (categorical variable $\times$ stereotype change $\times$ time) RM-ANCOVA controlling for gender, race/ethnicity, course section, and past science class experience.

Analysis of Student Grades
Students' course grades, expressed numerically (“A” = 4, “B” = 3, etc.), were included in analyses to explore correlations between Stereotypes, Nonstereotypes, relatability, and in-class achievement. As in tests for correlations involving interest in science, we used the categorical variables we generated for changes in Stereotypes, Nonstereotypes, and relatability in ANCOVAs to explore connections between those variables and course grades. These analyses controlled for gender, race/ethnicity, course section, and past science class experience.

All statistical analyses were performed in SPSS (SPSS for Windows, 19.0.0, IBM, Armonk, NY). To enhance clarity and readability, we present descriptive statistics and ANCOVA tables from our analyses in the Supplemental Material, parts E and F, rather than in the body of the article.

RESULTS
Hypothesis 1 Results: Scientist Spotlights Will Shift Students’ Descriptions of Scientists toward Nonstereotypes
Students’ weekly Scientist Spotlight responses suggested the assignments encouraged students to reflect on counterstereotypical examples of scientists while engaging with course content. Fernanda commented on her previous stereotypical ideas about scientists and discussed how Charles Limb counteracted those stereotypes by showing an interest in music and a life outside of science could contribute to a scientific career:

I was able to see scientists in a different perspective … I used to think scientists were mere geniuses who asked infinite, even unpredictable questions nobody had the time to research. I used to even think they were mere robots who ate, researched, and slept on a daily basis. Yet, they have a life of their own … I can tell Dr. Limb is a good musician whose love for the music stretched to his eagerness to learn about the brain.—Fernanda, a Latina student responding to the Scientist Spotlight on Charles Limb

Melissa noted that Raymond Dubois’s “humble beginnings” in an economically disadvantaged farming community represented a nontraditional path to science:

Dr. Dubois is such a unique person. He was born and raised to be a farmer, and didn’t have very much money or aspiration … He found science completely by accident and fell in love, and from such humble beginnings he became one of the country’s foremost experts in his field. It’s very impressive to see someone come from so traditionally unlikely a background and become so well-known for his work.—Melissa, a white female student responding to the Scientist Spotlight on Raymond Dubois

Shi...
students, as exemplified by Theresa’s response presented earlier. Many students resorted to describing scientists simply as those individuals who participate in certain, named scientific fields or professions. For example,

The types of people who do science are teachers, professors, NASA workers, nurses, doctors, etc. NASA scientists use science to study space and the earth … Doctors use science to study the human body.—Carlos, a Latino Course Reader Homework student

By the end of the course, most students from Scientist Spotlight classes used Nonstereotypes to describe scientists (Table 1A). Tania reflected on the ways her views of scientists changed and stated that many scientists defy stereotypes of individuals in their fields. Rather, scientists are “normal people” like her:

Before I learned about scientists in this class, I thought scientists were like “nerds” or what they show in movies. The characters would be very geeky, had glasses, spoke monotone, and thought they were above everyone. However, through all the research I’ve done in this class, scientists are just normal people like myself. They love to learn new things, they have a life outside the laboratory, they are fun … My opinion of people that do science has completely changed thanks to this class.—Tania, a Filipina Scientist Spotlight Homework student

Felipe reported that people from diverse countries and socioeconomic backgrounds are scientists and that scientists did not all have an innate interest in the field from an early age:

The types of people that do science are all kinds of people. Before I learned about scientists in this class, I thought scientists were like “nerds” or what they show in movies. The characters would be very geeky, had glasses, spoke monotone, and thought they were above everyone. However, through all the research I’ve done in this class, scientists are just normal people like myself. They love to learn new things, they have a life outside the laboratory, they are fun … My opinion of people that do science has completely changed thanks to this class.—Felipe, a student from Scientist Spotlight classes

### TABLE 1. Most common student descriptions of scientists from Scientist Spotlight Homework (A) and Course Reader Homework (B) students at the beginning and end of the course

| A. Beginning of Scientist Spotlight courses | Prevalence | End of Scientist Spotlight courses | Prevalence |
| People that do experiments (s) | 24% | All types of people (n) | 49% |
| Curious (s) | 20% | Passionate (s) | 24% |
| Biologists (f) | 18% | Cheerleaders (n) | 22% |
| Especially intelligent (s) | 17% | Darlene Cavalier (n) | 18% |
| Albert Einstein (s) | 17% | People that do experiments (s) | 17% |
| Doctors (f) | 15% | People from outside the United States (n) | 16% |
| People that look for “truths” (s) | 13% | Curious (s) | 15% |
| Chemists (f) | 13% | Creative (s) | 13% |
| Discover things (s) | 12% | Dedicated (s) | 13% |
| People that investigate natural world (s) | 11% | Interested in science (s) | 13% |
| Make the world better (s) | 11% | Go against stereotypes (n) | 12% |
| Enjoy learning (s) | 11% | Discover things (s) | 12% |
| Question things (s) | 10% | Rosalind Franklin (n) | 11% |
| Psychologists (f) | 10% | Make the world better (s) | 11% |
| Physicists (f) | 10% | Not just one type of person (n) | 11% |

| B. Beginning of Course Reader Homework course | Prevalence | End of Course Reader Homework course | Prevalence |
| Curious (s) | 32% | Curious (s) | 24% |
| Especially Intelligent (s) | 18% | Especially intelligent (s) | 15% |
| People that do experiments (s) | 14% | People that do experiments (s) | 14% |
| Discover things (s) | 13% | Discover things (s) | 14% |
| Interested in science (s) | 13% | All types of people (n) | 13% |
| Enjoy learning (s) | 12% | Chemists (f) | 12% |
| Albert Einstein (s) | 10% | Make the world better (s) | 12% |
| Chemists (f) | 10% | Doctors (f) | 11% |
| Doctors (f) | 10% | Biologists (f) | 11% |
| All types of people (n) | 8% | Albert Einstein (s) | 8% |
| People that investigate natural world (s) | 8% | Enjoy learning (s) | 8% |
| Open-minded (s) | 8% | People that investigate natural world (s) | 8% |
| Biologists (f) | 7% | Passionate (s) | 8% |
| Make the world better (s) | 7% | Creative (s) | 7% |
| Astronomers (f) | 7% | Geneticists (f) | 7% |

Shading and letters in parentheses denote categories of descriptions per Schinske et al., 2015: s/turquoise = Stereotype; n/light green = Nonstereotype; f/gray = Field of Science.
being from a different country to having a stereotypical assumption about a person, for example a cheerleader. Anyone can be a scientist if they want to. One thing all scientists we learned about had in common was that they weren’t interested in science until something sparked their interest.—Felipe, a Latino Scientist Spotlight Homework student

Matthew agreed that scientists need not be initially interested in science, citing the example of Carl Djerassi:

The types of people that do science vary greatly. One scientist, Djerassi, in an interview said he had no interest in science as a kid, but he eventually grew up to be the scientist that created contraceptive pills for women.—Matthew, a Vietnamese male Scientist Spotlight Homework student

Maria more specifically called attention to the fact that race and sex are not determinants of an ability to be a scientist:

All types of people can do science … What I learned was that your background/sex/race doesn’t determine if you will become a scientist or not. It is all about the passion and love for knowledge that human beings have.—Maria, a Latina Scientist Spotlight Homework student

Cynthia, as well as Tania (noted earlier), pointed out that interests outside of science can be as important to scientists as an interest in science:

[Scientists] take their passion and often combine it with science. For example, the scientist that was looking at musician’s [sic] brains as they improvised music.—Cynthia, a white female Scientist Spotlight Homework student

The above responses made the argument that many different types of people, and perhaps all types of people, are scientists. Indeed, at the end of the course, the majority of students (55%) included descriptions of scientists fitting into at least one of the following categories: all types of people, not just one type of person, or go against stereotypes. The quotations from Cynthia and Matthew further demonstrated that, at the end of the course, many students had specific, counter-stereotypical individuals in mind to inform their descriptions of scientists.

Matthew and Felipe pointed out that many scientists did not have an innate or early interest in science, and we no longer see references to scientists as especially intelligent in these exemplars. Given that we believe all of the scientists featured in Scientist Spotlights are very intelligent, we found it striking that scientists may still have been impressively smart, “intelligent” largely disappeared as ways to describe scientists (Table 1A). It appears that, while the featured scientists may still have been impressively smart, “intelligent” was no longer a significant defining feature of scientists in students’ minds. Rather, scientists were considered regular/normal people who happened to find their way to careers in science (responses of Matthew, Felipe, and Tania).

In contrast to the above findings from Scientist Spotlight students, Course Reader Homework students largely continued to use stereotypes and generalities to describe scientists at the end of the course (Table 1B). For example, Laila and Mei continued to describe scientists in terms of their special intelligence/knowledge:

People who work in science fields have absolutely incredible intelligence.—Laila, Indonesian female Course Reader Homework student

Scientists have to be up-to-date about research, medicine, diseases.—Mei, a Chinese female Course Reader Homework student

Carlos, like many other students in Course Reader Homework classes, continued to define scientists in nebulous terms through their fields/professions:

The types of people that do science are people that do astrophysics, astronomy, chemistry, biology, physics, and geophysical science. There are NASA scientists that study space. Also there are scientists that study humans and their environment.—Carlos, a Latino Course Reader Homework student

Theresa reiterated the importance of curiosity from her beginning-of-course response:

All kinds of people do science, especially those who are really curious about a certain scientific topic. Men can be scientists as well as women … Albert Einstein is a very famous scientist.—Theresa, a white female Course Reader Homework student

Theresa and some other Course Reader Homework students did mention at the end of the course that all types of people do science, causing that description to increase in prevalence (Table 1B). It is interesting to note, however, that the remainder of Theresa’s end-of-course response was nearly identical to her beginning-of-course response—emphasizing curiosity and raising the same example of Albert Einstein. In other words, while a small number of Course Reader Homework students appear by the end of the course to be describing a more inclusive version of who does science, those students’ responses still lacked the specific examples and expanded descriptions of scientists we observed from Scientist Spotlight students.

In quantitatively analyzing these trends, an RM-ANCOVA revealed significant interactions between treatment and the use of Stereotypes, $F(1,311) = 13.39$, $p < 0.001$, $\eta^2 = 0.04$, and Nonstereotypes, $F(1,311) = 16.51$, $p < 0.001$, $\eta^2 = 0.05$. When looking solely at raw means, we observed all students using fewer Stereotypes at posttest, but Scientist Spotlight Homework students showed a sharper decrease, suggesting that the treatment produced a stronger decrease in Stereotype use. However, an analysis of weighted means to isolate the variability introduced by treatment condition from the variability introduced by race/ethnicity, gender, and course section, showed no significant differences in the decrease across groups. In terms of Nonstereotypes, both raw and weighted means show a significant increase among Scientist Spotlight students when compared with Course Reader Homework students (Figure 1 and Supplemental Material, parts E and F). Therefore, when controlling for unequal group sizes and nonrandom assignment,
our results suggested the completion of Scientist Spotlights was associated with increases in the use of Nonstereotypes in describing scientists.

Hypothesis 2 Results: Scientist Spotlights Will Enhance Students’ Ability to Personally Relate to Scientists

Scientist Spotlight Homework submissions provided evidence of students encountering scientists to whom they could relate on a personal level. For example, Binh could relate to Flossie Wong-Staal and Juan Perilla because, like him, they were originally from outside the United States, albeit from countries different from his:

Another thing is scientists who are successful in the U.S. are not necessary [sic] born in the U.S. These scientists are both from another country but they’re really successful. It makes me more confident in becoming a scientist because no one in my family is a scientist and I’m not a U.S. citizen.—Binh, a Vietnamese male student responding to the Scientist Spotlight on Flossie Wong-Staal and Juan Perilla

On the other hand, Emily could relate to Charles Limb due to shared interests outside science:

I found this Ted Talk with Charles Limb incredibly interesting mostly because I am a musician myself who has been trained both classically and in jazz.—Emily, a white female student responding to the Scientist Spotlight on Charles Limb

Anthony found Agnes Day relatable due to their shared racial/ethnic identities and because of what she represents to people like him:

For my whole life I … wasn’t exposed to any scientist who was of African American descent. That, as a fellow African American, brought me joy as it shows that African Americans are no longer abiding to the negative stigma we have. She’s representing a powerful position for us and people have noticed her work. It gave me incentive to push for my own dreams and to succeed.—Anthony, a Black male student responding to the Scientist Spotlight on Agnes Day

Some of the resources students reviewed during Scientist Spotlights demonstrated that scientists experienced barriers, inequities, and marginalization or that science itself can include the study of social inequities (e.g., health disparities). These themes spurred many students, like Anthony, to connect with scientists through the lens of social justice. After learning about Ben Barres’s personal story and path in science, Maria discussed her views on gender equity in science and how that relates to her experience at her community college. She further compares what she learned about the biology content in this assignment (glial cells) with the plight of women in science:

The fact that there are considerably less women in science than men, is more of a socio-cultural problem, than a genetic or gender problem. Personally, I feel optimistic, yes we are the minority in science, and are paid less than men, and are discriminated against, but when I look around my community college I see many women succeeding, and unafraid to give the best of themselves … In a way glia cells are a little bit like the “women” of the nervous system; extremely important for the survival of the cells, form the majority of the nerve cells population, and are underestimated and perceived only as a “supporter” cell.—Maria, a Latina student responding to the Scientist Spotlight on Ben Barres

Gina responded to Agnes Day’s scientific work by proposing that the type of science that gets done might depend largely on the type of people doing the science. As a result, diversity in the sciences might be required in order to understand the importance of, and go on to pursue, certain research areas:

Dr. Day is one of the first to complete a study in cancer concerning the differences in race. If she was not African American I do not think that Dr. Day would understand the significance of her research … As a strong Black woman representing women and people of color in a White male driven field Dr. Day defies what I believed about people who do science. I wonder if the questions of science require diversity, collaboration and personal passions in order to be answered.—Gina, a Black/Native American female student responding to the Scientist Spotlight on Agnes Day

Beginning- and end-of-course responses to the relatability prompt additionally demonstrated distinct shifts in an ability to personally relate to scientists. Two hundred eight Scientist Spotlight Homework students and 86 Course Reader Homework students submitted both beginning- and end-of-course responses to the relatability prompt. The sample size for this prompt was smaller than that for the stereotypes prompt, since it took longer to develop and establish face validity for this prompt. As a result, it was only presented at both time points to four of the five sections of Scientist Spotlight students. The final relatability prompt stated: “I know of one or more important scientist to whom I can personally relate,” which was followed by a Likert scale and a space for qualitatively explaining the opinion selected. An “I don’t know” option was included in the Likert scale and was
coded as “Disagree” based on the qualitative explanations provided by students selecting “I don’t know” (e.g., “I honestly only know of one [scientist] and I’m nothing like him”).

Only 35% of students in Scientist Spotlight Homework classes and 36% in the Course Reader Homework classes either agreed or somewhat agreed with the relatability prompt at the start of the course, indicating that students did not generally feel they could relate to scientists. Students’ beginning-of-course responses regarding their ability to relate to scientists fell into two main categories. First, as exemplified by the responses of Jesus and Evelyn, many students explicitly affirmed that they were unable to relate to scientists:

I Don’t Know. I truly am terrible at relating to people that are involved with science or math.—Jesus, a Latino Scientist Spotlight Homework student

Disagree. I don’t personally relate to any scientist as most of my friends and family members are not scientists.—Evelyn, a Chinese female Course Reader Homework student

Ademar and Beth clarified that this was often because students lacked familiarity with any actual scientists:

Disagree. I personally don’t know any scientist, and sometimes I cannot see myself having the personal qualities of a scientist.—Ademar, a Latino Course Reader Homework student

I Don’t Know. I’m not very familiar with scientists or their names and studies.—Beth, a Black/Latina female Course Reader Homework student

Second, among the few students who indicated at the beginning of the course they could personally relate to scientists, many, like Yvette, explained this was simply because they appreciated the types of work scientists did:

Somewhat Agree. I am knowledgeable of various scientists but I don’t feel personally relatable to them. I appreciate their work and what it has done to better inform us as a society.—Yvette, a Latina Scientist Spotlight Homework student

At end of the course, 79% of Scientist Spotlight Homework students agreed or somewhat agreed that they could personally relate to an important scientist. These students’ end-of-course explanations differed markedly from their beginning-of-course responses and included many details as evidence for relating to (or not relating to) scientists. Two main themes arose as reasons students related to scientists at the end of the course. First, many students found they could relate to scientists due to shared interests or personal qualities. Lauren described how she could relate to Charles Limb due to common interests surrounding music:

Agree. I relate the most with the neurologist/musician from the first scientist spotlight … because I am also a musician.—Lauren, a white female Scientist Spotlight Homework student

Jesus, on the other hand, related to Lawrence David due to a shared sense of humor, an interest in making others laugh, and a similar work ethic:

Agree. I can relate to that one scientist who interacted with poop. I loved his sense of humor and drive to complete an experiment … I know that I can relate to him because I love being funny to make people smile and also am determined to work on things until I finish.—Jesus, a Latino Scientist Spotlight Homework student

Second, some students found scientists relatable if the scientists did not originally expect to enter a career in science. Yvette found she could relate to many of the scientists for this reason and further explains that she is similarly reconsidering her interest in studying science:

While a less common theme, seeing scientists with matching genders or races/ethnicities was important in making them relatable for some students, like Rachel:

Somewhat Agree. In some of the spotlights some scientists felt that they didn’t always want to pursue a career in science and that it just happens. I’m starting to feel the same way. I’m not originally a science major but I feel that I could have a future in it if I find the right field.—Yvette, a Latina Scientist Spotlight Homework student

Others, like Tammy, indicated that it made scientists more relatable to see they have encountered similar struggles or injustices in life:

Somewhat Agree. Although I might not be that interested in pursuing a career in science, being exposed to a wide variety of diverse scientists, I feel like I could go into this field if I wanted to. Many of the scientists we learned about were women and many were a race other than White. These are both characteristics I would use to describe myself.—Rachel, a Filipina Scientist Spotlight Homework student

As seen in earlier quotes, many students at the end of the course were able to name or describe specific scientists in their responses, suggesting greater familiarity. Of course, this familiarity did not always result in relatability. Amit simply could not envision himself having the same passion for science:

Disagree. In our scientist spotlights, all the scientists came from very different backgrounds. However, they all liked science very much. I can’t relate to that. I don’t have any particular disdain for science, but I don’t enjoy it. I do think it is very important, however.—Amit, an Asian Indian male Scientist Spotlight Homework student

This presented a barrier to finding scientists relatable, even when recognizing the featured scientists were very diverse. On the other hand, notable shifts in qualitative responses toward an increased ability to relate to scientists were sometimes observed even among students whose Likert-scale relatability
selections did not change (e.g., Yvette, who selected “somewhat agree” at both the beginning and end of the course).

Only 43% of Course Reader Homework students agreed or somewhat agreed with the relatability prompt at the end of the course. End-of-course qualitative responses from these students were strikingly similar to their beginning-of-course responses, with many students, like Evelyn and Beth, using language identical to what they had written at the beginning of the course:

I Don’t Know. None of my friends or family members are scientists.—Evelyn, a Chinese female Course Reader Homework student

Somewhat Disagree. I am not very familiar with scientists.—Beth, a Black/Latina female Course Reader Homework student

Responses reiterated beginning-of-course themes that most students could not relate to, and did not even know of, any scientists. This was in spite of the fact that some scientists were introduced as part of certain lectures during Course Reader Homework classes (see Methods).

Following an RM-ANCOVA, we observed an interaction between treatment × time for relatability Likert-scale ratings on the relatability prompt, $F(1,276) = 8.49, p = 0.004, \eta^2 = 0.03$. Course Reader Homework students’ end-of-course relatability Likert scores did not differ significantly from their beginning-of-course scores, while Scientist Spotlight students’ end-of-course relatability scores were significantly higher than both their own beginning-of-course scores and Course Reader Homework participants’ end-of-course scores (Figure 2 and Supplemental Material, parts E and F). Quantitative results therefore support the hypothesis that Scientist Spotlights increase students’ sense of relating to scientists.

Evidence Regarding Longitudinal Impacts of Scientist Spotlights on Stereotypes and Relatability
Fifty-seven Scientist Spotlight Homework students submitted a response to the stereotypes prompt 6 months after the end of their courses (17% response rate). Of those, 47 had submitted responses to the stereotypes prompt at all three time points (beginning of term, end of term, 6 months after class). Fifty-two students submitted a response to the relatability prompt 6 months after the end of their courses (15% response rate). Of those, 27 had submitted responses to the relatability prompt at all three time points. As the community college student population is in constant flux, with students transferring to 4-year schools or professional programs, moving between colleges, and entering and exiting school at various times due to work and family obligations, we were not surprised by the modest response rate to a survey 6 months after the end of class. In spite of these lower sample sizes, however, this 6-month follow-up subsample appeared to match the larger sample in terms of demographics. Three independent $t$ tests for gender, race/ethnicity (traditionally underserved vs. traditionally well served), and condition demonstrated that gender, $t(279) = −0.655, p = 0.513$, and race/ethnicity, $t(69.87) = 0.908, p = 0.367$, were similar between the 6-month follow-up sample and the larger, original sample.

Six months after the end of class, students appear to have maintained the largely nonstereotypical ideas about scientists they displayed at the end of the course. Table 2 shows the most prevalent themes found in responses to the stereotypes prompt from students who submitted essays at all three time points. We additionally created word clouds to visually convey the full range of scientist descriptions at each time point (Supplemental Material, part I). Descriptions of scientists as representing many/all types of people remained the most common theme in the 6-month postclass responses. Students additionally continued to describe scientists as individuals who defy stereotypes, and the idea that scientists have “special intelligence” continued to be relatively rare. Fifty-seven percent of students included descriptions of scientists fitting into at least one of the following categories 6 months after the course: all types of people, not just one type of person, and go against stereotypes.

Three-way RM-ANCOVAs controlling for gender and race/ethnicity (Supplemental Material, parts E and F) showed that stereotypical descriptions dropped significantly at the end of the course and remained low 6 months later, $F(2,78) = 4.36, p = 0.016, \eta^2 = 0.10$ (Figure 3a). Nonstereotypical descriptions increased significantly at the end of the course and remained high 6 months later, $F(2,80) = 5.97, p = 0.004, \eta^2 = 0.13$ (Figure 3b). Relatability similarly increased at the end of the course and remained high 6 months later, though in this case the initial increase was detected at a $p$ value of 0.083, $F(2,46) = 2.63, p = 0.083, \eta^2 = 0.10$ (Figure 3c). This was likely because of the smaller sample size available for the relatability prompt.

Hypothesis 3: Shifts in Scientist Stereotypes and Relatability of Scientists Will Correlate with Students’ Interest in Science
We calculated both beginning- and end-of-course Science Interest scores (Supplemental Material, parts G and H) for each student. To test the relationship between shifts in Science Interest and shifts toward majoring in STEM fields, we conducted a 2 × 2 (Science Interest × STEM major interest) RM-ANCOVA controlling for gender, race/ethnicity, course section, and prior science class experience. Values for STEM major interest came from the online survey item “I am majoring or plan on majoring in another Science or Math field” (Supplemental Material, part D).
A significant interaction for Science Interest was found, $F(1,216)=10.39$, $p = 0.001$, $\eta^2 = 0.05$, in which students whose Science Interest decreased or held steady showed a significant decrease in STEM major interest from pretest ($\bar{x} = 3.70$, SE = 0.16) to posttest ($\bar{x} = 3.43$, SE = 0.18), while students whose Science Interest increased reported more STEM major interest at posttest ($\bar{x} = 3.34$, SE = 0.16) than at pretest ($\bar{x} = 3.74$, SE = 0.18).

RM-ANCOVAs using the Science Interest scale (Supplemental Material, parts E and F) revealed that a decrease in the use of Stereotypes correlated with higher Science Interest at the end of the course, $F(1,182) = 4.46$, $p = 0.036$, $\eta^2 = 0.02$ (Figure 4a). We found a similar relationship between an increase in the use of Nonstereotypes and Science Interest that approached significance, $F(1,182) = 3.32$, $p = 0.070$, $\eta^2 = 0.02$ (Figure 4b). Science Interest additionally appeared to increase from beginning of course ($\bar{x} = 3.287$, SE = 0.076) to end of course ($\bar{x} = 3.568$, SE = 0.061) for students whose ability to relate to scientists increased, but this finding did not achieve statistical significance, $F(1,184) = 2.10$, $p = 0.149$, $\eta^2 = 0.01$. In total, these results provide partial support for the hypothesized relationship between shifts in scientist stereotypes/relatability and an interest in science/STEM majors.

**Hypothesis 4:** Shifts in Scientist Stereotypes and Relatability of Scientists Will Correlate with Course Grades

As a first step, we tested whether the treatment had an effect on course grades. A one-way ANCOVA, controlling for gender, race/ethnicity, course section, and previous science class experience, revealed that Scientist Spotlight Homework students earned significantly higher grades than Course Reader Homework students, $F(1,279) = 6.68$, $p = 0.018$, $\eta^2 = 0.02$ (Figure 5a and Supplemental Material, parts E and F).

Additional analyses were limited to Scientist Spotlight Homework students to prevent confounds introduced by the treatment. One-way ANCOVAs suggested there was not a significant effect for the use of Stereotypes on grades, $F(1,211) = 3.00$, $p = 0.085$, $\eta^2 = 0.01$, but there was a significant effect of Nonstereotypes, $F(1,211) = 6.68$, $p = 0.010$, $\eta^2 = 0.03$. Students whose use of Nonstereotypes increased earned significantly higher course grades than those whose use of Nonstereotypes held steady or decreased (Figure 5b and Supplemental Material, parts E and F). To test the relationship between relatability and course grade, we compared students whose relatability ratings increased, those whose relatability ratings decreased, and those whose ratings held steady. A one-way ANCOVA controlling for race/ethnicity, gender, course section, and science experience, suggested the grades of students whose ratings decreased ($\bar{x} = 2.59$, SE = 0.24) were lower than students whose ratings held steady ($\bar{x} = 2.79$, SE = 0.15) or increased ($\bar{x} = 3.01$, SE = 0.10). However, the difference between groups was not significant, $F(1,171) = 1.65$, $p = 0.195$, $\eta^2 = 0.02$. The finding of a correlation between an increase in Nonstereotypes and course grades therefore provided partial support for hypothesis 4.

**DISCUSSION**

Many reports have documented the shortfall in students graduating with STEM degrees in the United States and the urgent need to recruit a more diverse STEM workforce (National Academy of Sciences, 2007, 2011). Interventions with the potential to enhance students’ science identities and reduce stereotype threat could prove valuable in promoting interest and success in STEM (Seymour and Hewitt, 1997; Brickhouse et al., 2000; Hill et al., 2010, chap. 3; Beasley and Fischer, 2012). We developed and tested an intervention in the form of weekly homework assignments that were aimed at allowing students to see their possible selves in science and promoting counterstereotypical examples of who does science. In the following sections, we discuss the utility of Scientist Spotlights in light of our...
Intervention to Enhance Science Identity

Scientist Spotlights Generated Shifts in Students’ Stereotypes of Scientists and Scientist Relatability

We used the stereotypes prompt to evaluate the impact of Scientist Spotlights on students' stereotypes of scientists. When compared with a class performing a similar activity that lacked connections with diverse scientists, students who completed Scientist Spotlights adopted more nonstereotypical views of scientists (Figure 1). These changes appeared to be sustained 6 months after the courses ended (Figure 3) and were associated with higher course grades (Figure 5). Reductions in stereotypical descriptions of scientists further correlated with increases in Science Interest (Figure 4a) and an enhanced interest in STEM majors.

We piloted the relatability prompt as a tool for examining students' possible selves in a science context, making the case that explicitly asking students about their ability to personally relate to scientists would draw out descriptions of students' possible selves in relation to scientists. While only 43% of Course Reader Homework students found scientists relatable at the end of the course, the vast majority (79%) of Scientist Spotlight students did (Figures 2 and 4c). These students discussed shared personalities and interests outside science as reasons for being able to relate to scientists, with some students also commenting on certain scientists' nontraditional paths to gaining an interest in science. Many students used specific language such

![FIGURE 3](image1.png)

![FIGURE 4](image2.png)
as “like me” or “I am also ...” when describing why common interests or personal qualities caused them to relate to scientists after Scientist Spotlights. This suggested the relatability prompt might have functioned as intended in creating opportunities for students to reflect on their possible science selves.

These findings suggest Scientist Spotlights hold promise as a tool for enhancing students’ possible science selves and disrupting stereotypes of scientists in diverse classroom settings. Prior studies point to the importance of these shifts in forming a science identity, mitigating stereotype threat, and enhancing student interest and success (Steele, 1997; Oyserman et al., 2006; Steinke et al., 2009; Hill et al., 2010, chap. 3; Hunter, 2010; Beasley and Fischer, 2012; Mills, 2014).

Scientist Spotlights Represent a Simple Means for Raising Issues of Diversity in STEM Classrooms

Faculty might feel particularly wary of adopting new activities that overtly approach issues related to race and diversity due to a lack of training in how to facilitate discussions in those areas (Sue et al., 2009). STEM faculty commonly cite course content expectations and concerns regarding time as barriers to implementing innovative teaching strategies (Henderson and Dancy, 2007; Austin, 2011). Scientist Spotlights offer faculty an approach for openly addressing diversity in STEM classes while supporting content goals and requiring little grading or class time.

Because Scientist Spotlights are assigned as homework and are graded based on timeliness and word count, the activities consume only a negligible amount of instructor time during and outside of class. This is perhaps particularly the case when they are assigned through an online course management system that automatically displays word counts. After an initial investment of time to identify scientists to feature and compose assignment prompts, Scientist Spotlights become an easily sustainable class activity.

Additionally, by connecting diversity themes to course content through Scientist Spotlights, faculty are able to structure some of students’ content learning outside class. In this way, Scientist Spotlights assist faculty in meeting their content expectations, rather than taking time away from addressing content. This follows the best practices discussed by Chamany et al. (2008), who recommend “strategically embedding social context into those topics that are traditionally reviewed in ... biology courses.” Highlighting the struggles and inequities experienced by scientists like Ben Barres also opened up opportunities for students to engage with issues of social justice in science. Infusing course content with themes of equity and social justice has been promoted as a particularly impactful way to engage traditionally underserved and underprivileged populations of students in STEM (Chamany, 2006; Chamany et al., 2008). At the same time, these themes of equity and diversity were clearly contextualized within instructors’ comfort zone of course content, which might allay instructor reservations about raising such themes as part of a STEM class.

We predict that the strongest case for faculty adoption of Scientist Spotlights, and eventually adoption of more extensive diversity-related activities, might come from students themselves once faculty pilot Scientist Spotlights. Students in our sample responded so immediately and effusively to Scientist Spotlights, it appeared there was a great, unmet demand among students to approach science content through this lens. We predict that, if faculty see responses from their own students similar to those shown here, they will feel energized and empowered to become more deeply involved in addressing diversity. Scientist Spotlights might therefore represent an excellent introductory tool that could inspire further work on equity and diversity in STEM by science faculty.

Suggestions for Implementation

While Scientist Spotlights are relatively simple activities, successfully implementing them in a course likely depends in part on how an instructor chooses scientists to feature, writes the assignment prompts, introduces the assignments to the class, and reports back on students’ submissions. In the following sections, we elucidate some of the factors we feel assisted in achieving positive outcomes and reducing the potential for student resistance.

Possible Selves as a Framework for Selecting Scientists to Feature in Spotlights

We found the concept of possible selves to be helpful in identifying scientists to feature. Rather than looking for scientists to serve as role models that students should emulate, we sought out scientists with whom students might already have similarities; that is, scientists in whom students might see their possible selves. While gender/race/ethnic matching was important for
some students, students more often cited shared personal qualities and outside interests as ways in which they saw themselves in scientists. Given that Human Biology primarily serves non–biology majors, it is not surprising that students also appreciated that not all scientists aspired to a science career at a young age and sometimes found science later in life. In consideration of the above, it is important to identify scientists for whom some sort of engaging biographical resource exists. It was in those biographical resources that students most directly encountered counterstereotypical information about scientists and found information that reminded students of themselves. We optimally hoped to find TED Talks, interviews, or podcasts featuring scientists telling their own stories in their own voices. However, we sometimes used printed interviews and biographical information, as in the example regarding Ben Barres (see Methods).

The Story Collider (www.storycollider.org/podcasts) proved a particularly rich resource for identifying biographical information regarding counterstereotypical scientists. The Story Collider website includes hundreds of 10- to 20-min-long, often funny or emotionally stirring autobiographical stories told by diverse scientists. The podcast descriptions can be searched for certain key terms through the website, which can be helpful in identifying scientists working in areas connected with course content.

**Metacognition as a Design Feature of Scientist Spotlight Prompts**

In terms of the assignment prompt itself and the regularity of the assignments, our work suggests that performing Scientist Spotlights regularly and including a metacognitive question about who does science assisted in achieving the outcomes we observed. Course Reader Homework classes included three references to scientists working in the fields being studied in class (see Methods). Two of those scientists identified as people of color and all three had counterstereotypical qualities. Students were introduced to those scientists during class, saw pictures of the scientists, and watched short videos featuring two of the scientists. However, students did not engage in any individual or group activities regarding the scientists and were not asked to reflect on whether those segments of class impacted their views of scientists. Our results suggested these students did not substantially change their views of scientists. This suggests that going beyond simply mentioning/showing diverse scientists in class and moving to require regular work including metacognition about who does science might be key for stimulating larger changes in the ways students view scientists. Science faculty are increasingly aware that metacognition is necessary to drive lasting changes in students' ideas and behaviors (Tanner, 2012). We therefore propose that the prompt reading, “What do these resources tell you about the people that do science?”, might be important to include in every Scientist Spotlight assignment, even if the other writing prompts vary from one assignment to the next.

**Instructor Talk as a Strategy for Securing Student Buy-In**

Alongside content expectations and time limitations, fear of student resistance represents another of the main barriers to the adoption of new teaching strategies by faculty (Henderson and Dancy, 2007; Seidel and Tanner, 2013). We encountered very little evidence of student resistance to completing Scientist Spotlights in these classes. Students completed Scientist Spotlights at very high rates, earned high scores, and seemed to find the assignments engaging and helpful. Students’ acceptance of Scientist Spotlights might partially relate to the flexibility students had to engage with either the course content part of the activity or the scientist biography part of the activity. Students were allowed to independently determine how much of their submissions focused on the “types of people that do science” prompt compared with the course content--related prompts. In this way, students could settle into their own comfort zones of discussing issues of content versus issues of diversity and scientist stereotypes.

The non–content language instructors use to frame new activities and debrief completed activities (“instructor talk”) might additionally play a large role in reducing student resistance and creating effective environments for applying innovative strategies (Seidel et al., 2015). While Scientist Spotlights are largely out-of-class activities, J.N.S. spent a small amount of class time at the start of the course establishing a classroom culture conducive to performing Scientist Spotlights and explaining his pedagogical decision to use these assignments. Specifically, he made clear his reasons for incorporating Scientist Spotlights into the course and his goals for the assignments, expressed that there were no “right” or “wrong” ways to respond, and noted that students could write about whatever parts of the assignments resonated most strongly with them each week. They need not strictly respond to each assignment prompt in equal amounts or in the order shown.

Following the first and second Spotlights, J.N.S. spent ~5 minutes in class sharing anonymous student quotes to demonstrate how different students engaged with course content and reflected on their notions of scientists through the assignments. J.N.S. especially looked for quotes similar to Gina’s (discussed earlier) demonstrating the importance of the types of people who do science to the types of scientific questions that get pursued. This showed students in their own words that diversity is necessary to ensure diverse scientific questions are addressed and that it is important to understand who does science when considering what currently is and is not known about the topics studied in class.

**Limitations**

While quasi-experimental studies can represent a robust means of addressing education research questions, it is critical to explore alternate explanations for outcomes that might stem from the lack of random assignment to quasi-experimental groups (Shadish et al., 2002). Though the course sections we studied were equivalent in many respects, they differed slightly in student demographics, timing during the year, and lecture location. It is possible, for example, that differences observed between Scientist Spotlight Homework and Course Reader Homework groups were influenced by slight variations in student racial/ethnic or gender identities between those groups. This would confound our ability to attribute differences to our intervention. Similar scenarios could be proposed for differences in lecture locations or timing during the year. However, all lecture rooms were similarly appointed and neither treatment group was isolated to a single part of the year. The five Scientist Spotlight courses took place throughout the year (three Fall classes, one Winter class, one Spring class), while one Course Reader Homework class took place in the Fall and the other in the Spring.
Though differences between the courses appeared relatively subtle, we used statistical corrections to partition out variance introduced by demographics, course section differences, and the unequal sizes of quasi-experimental groups (i.e., lower number of Course Reader Homework students). The resulting “weighted means” were used in evaluating our hypotheses. These weighted means often differed substantially from means observed in our raw data (Supplemental Material, part E). This provided us more assurance that the differences we observed were due to the Scientist Spotlights but at the cost of variability that may have demonstrated a more robust effect. As a result, it might be argued that our results provide only conservative estimates of the impacts of Scientist Spotlights due to overly aggressive statistical corrections. That said, some researchers argue that statistical corrections are still insufficient to account for a lack of randomization, and issues with unequal group characteristics could confound the ability to make strong inferences (Shadish et al., 2002).

Other differences between our quasi-experimental groups included drop/fail/withdrawal (DFW) rates and the fact that one Course Reader Homework group was cotaught with a second instructor. From our results, it is apparent that 72% of Scientist Spotlight Homework students submitted both a beginning- and end-of-course stereotypes prompt essay, but only 67% of Course Reader Homework students did so. This might partially relate to differences in DFW rates between Scientist Spotlight and Course Reader Homework classes, effectively resulting in higher attrition in Course Reader Homework classes. Scientist Spotlight Homework classes had a 20% DFW rate compared with a 23% DFW rate in Course Reader Homework classes (for reference, the average DFW rate across all Human Biology classes at this college is 29%). It is also possible that Course Reader Homework students were less engaged in class, causing more of them to miss one of the days when a survey was scheduled. In either case, if the lower response rate among Course Reader Homework classes occurred disproportionately among students who shifted toward higher levels of Nonstereotypes/relatability, then attrition in those classes could partly account for differences observed between quasi-experimental groups. This scenario seems unlikely, however, given that our findings suggest students conveying higher levels of Nonstereotypes and relatability have increased success in class (Schinske et al., 2015; current study). It seems more likely that attrition could have masked larger differences between our groups by eliminating additional data points for Course Reader Homework students who did not shift in these variables.

It is also possible that the addition of a coteacher for one Course Reader Homework section influenced these differences between groups as well as our results. However, J.N.S. maintained control over relevant course assignments in all sections, and the cotaught section was equivalent to the others in terms of its curriculum expectations and types of class activities. Further, we included course section as a covariate in analyses to control for course-level differences. While we observed significant variation in dependent variables among students, we did not observe such variation between course section groups.

With regard to descriptions of scientists reported from student essays, our study did not seek to establish certain descriptions as “good” and others as “bad” in relation to enhancing success or interest in biology. While some studies have categorized certain scientist stereotypes as “positive” and “negative” (Mead and Metraux, 1957), we did not explore students’ cultural evaluations of specific stereotypes and cannot conclude whether individual students view such associations positively or negatively. Further surveys and interviews would be necessary to evaluate the deeper meanings and relative importance of various descriptions within the Stereotypes and Nonstereotypes categories. It should additionally be noted that our results do not provide specific insights regarding the mechanism(s) behind the outcomes observed surrounding Scientist Spotlights. Future work could explore the roles of metacognition, stereotype threat reduction, identification of possible selves, and other factors as mechanisms underlying these results.

Other possible limitations involve our proposed assessment of students’ possible science selves and the nature of our survey activities more generally. We used the concept of “relatability” as a means of capturing possible selves, making the case that the prompt explicitly asked students about whether they could relate to a scientist they knew. This was an exploratory narrative approach, and whether it fully captures a student’s sense of their own potential talents and abilities as scientists is a question for further exploration. Our measure was also limited in its ability to capture how students thought of themselves in terms of the characteristics of scientists they named. A more precise measure of students’ sense of self-as-scientist could be helpful to expand upon and clarify the present findings.

Finally, results presented in this paper might not be broadly generalizable to all school settings. Qualitative studies have the strength of more deeply exploring student ideas but can lack the generalizability of some quantitative studies (Johnson and Christensen, 2008, pp. 441–442). We conducted our study in the unique environment of a large, diverse community college in the San Francisco Bay Area. One might anticipate different results or student reactions in less diverse settings in different parts of the United States. The types of exemplar quotes we report and the frequencies of themes we observed in students’ essays, therefore, might be specific to our student population and teaching context.

**Future Directions**

We envision multiple opportunities to extend this work in the future, ranging from further explorations of the present findings in Human Biology classes to dissemination of the intervention across new institutions and teaching contexts. In light of the limitations discussed in the previous section, pursuing study designs that match students to quasi-experimental groups or randomize participants could reveal further significant trends and more fully illuminate the impacts of the intervention. Assessing Scientist Spotlights in additional class contexts would assist in exploring the generalizability of our findings. We also believe further explorations of the relatability prompt and other measures that might evaluate students’ possible science selves could yield valuable insights into broadening participation in STEM. For example, while we observed intriguing trends connecting shifts in relatability to broader student outcomes, such as higher Science Interest and course grades, these trends did not achieve statistical significance. Further studies of relatability would assist in more fully illuminating its connections to these broader outcomes and clarifying its relationship to the broader concept of possible science selves.
Future studies might additionally more directly explore the impacts of Scientist Spotlights on stereotype threat or classroom equity gaps. That certain shifts related to Scientist Spotlights correlated with increased Science Interest and higher course grades is encouraging and raises interesting questions about how students of different genders and races/ethnicities experienced these outcomes. However, our unequal group sizes and the nonrandom distribution of students among conditions prevented us from drawing conclusions along these lines. Further, the trends we observed in Science Interest were in relation to shifts in stereotypes/relatability, not treatment effects. Observing treatment effects related to Science Interest might require more robust controls and might be assisted by studies exploring students’ sense of themselves as scientists in relation to Science Interest. Additional longitudinal data would also assist in understanding the enduring impacts of Scientist Spotlights. Longer-term follow-up data from both Scientist Spotlight students and control students would allow us to investigate how sustained shifts in stereotypes and relatability correlate with motivation and behavior in the future, specifically as they relate to pursuing and persisting in STEM majors.

Perhaps the most exciting extension of this work involves engaging additional faculty in the creation and deployment of Scientist Spotlights in new institutional and classroom contexts. Through our workshops and presentations at conferences, a wide array of faculty from diverse STEM (and non-STEM) fields have expressed interest in using Spotlights in class. The only somewhat time-consuming step in using Scientist Spotlights is the work done before the start of a course to select scientists, gather appropriate scientific and biographical resources regarding the scientists, and compose the assignment prompts. It might therefore be useful to nucleate a community of STEM faculty to build Scientist Spotlight modules for many different curricular areas. This could result in a database of ready-to-use assignments matching a wide range of content areas and could additionally build a strong community of STEM educators focused on issues of equity and diversity.

ACKNOWLEDGMENTS
We extend our appreciation to Kimberly Tanner, Jennifer Myhre, the monitoring editor, and three anonymous reviewers for providing valuable feedback with regard to this article and to Jahana Kaliangara and Monica Cardenas for assisting in processing and presenting preliminary data leading up to this study. J.N.S. thanks Sonya Dreizler, Veronica Neal, Mallory Newell, IMPACT AAPI, and the Equity Action Council at De Anza College for their support. The organizers of the Conference on Understanding Interventions That Broaden Participation in Science Careers kindly provided travel funding to support our presentation of preliminary findings from this work in a lunchtime plenary in 2015. IMPACT AAPI and the Office of Staff and Organizational Development at De Anza College have generously provided J.N.S. and A.S. with travel funds to present on Scientist Spotlights at national meetings.

REFERENCES

Beardslee DC, O’Dowd DD (1961). The college-student image of the scientist: scientists are seen as intelligent and hard-working but also as uncultured and not interested in people. Science 133, 997–1001.


Martin D (2015). Women in science: are portrayals on primetime television negative, and what are effects of exposure to such content? Doctoral Dissertation, University of Missouri–Columbia.


New Measures Assessing Predictors of Academic Persistence for Historically Underrepresented Racial/Ethnic Undergraduates in Science

Angela Byars-Winston,†* Jenna Rogers,† Janet Branchaw,† Christine Pribbenow,† Ryan Hanke,* and Christine Pfund‡

†Department of Medicine, Center for Women’s Health Research, University of Wisconsin–Madison, Madison, WI 53715; ‡Center for Women’s Health Research, University of Wisconsin–Madison, Madison, WI 53715; §WISCIENCE, Wisconsin Institute for Science Education and Community Engagement, University of Wisconsin–Madison, Madison, WI 53715; ¶Wisconsin Center for Education Research, University of Wisconsin–Madison, Madison, WI 53715; †Department of Educational Psychology, University of Wisconsin–Madison, Madison, WI 53715; ‡Department of Medicine, Institute for Clinical and Translational Research, Center for Women’s Health Research, and Wisconsin Center for Education Research, University of Wisconsin–Madison, Madison, WI 53715

ABSTRACT
An important step in broadening participation of historically underrepresented (HU) racial/ethnic groups in the sciences is the creation of measures validated with these groups that will allow for greater confidence in the results of investigations into factors that predict their persistence. This study introduces new measures of theoretically derived factors emanating from social cognitive and social identity theories associated with persistence for HU racial/ethnic groups in science disciplines. The purpose of this study was to investigate: 1) the internal reliability and factor analyses for measures of research-related self-efficacy beliefs, sources of self-efficacy, outcome expectations, and science identity; and 2) potential group differences in responses to the measures, examining the main and interaction effects of gender and race/ethnicity. Survey data came from a national sample of 688 undergraduate students in science majors who were primarily black/African American and Hispanic/Latino/a with a 2:1 ratio of females to males. Analyses yielded acceptable validity statistics and race × gender group differences were observed in mean responses to several measures. Implications for broadening participation of HU groups in the sciences are discussed regarding future tests of predictive models of student persistence and training programs to consider cultural diversity factors in their design.

INTRODUCTION
There is a mismatch between the cultural demographics of the nation and the scientific workforce. This mismatch persists despite ongoing efforts to increase the participation of groups historically underrepresented (HU) in the sciences, including women and individuals from some racial and ethnic groups (Valantine and Collins, 2015). National efforts launched to diversify the scientific enterprise, ranging from the White House’s 2009 Educate to Innovate initiative to the America COMPETES Act of 2007 and 2010, have all been targeted at increasing participation in career pathways in science, technology, engineering, and mathematics (STEM) for HU groups. Yet the progress of these groups earning science degrees, choosing science career paths, and entering the scientific workforce has been minimal (National Science Foundation [NSF], 2013). This minimal participation stands in sharp contrast to women, African Americans, Latino/as, and Asian Americans comprising both the majority of the U.S. population and, consequently, the largest
share of new labor market entrants (Byars-Winston, 2014; Valantine and Collins, 2015).

In general, more than half of undergraduate first-year students from HU racial/ethnic groups, namely African Americans, Latino/as, and Native Americans, intend to major in a science or engineering degree (Hurtado et al., 2007; President’s Council of Advisors on Science and Technology, 2012). But together they earn only ∼17% of bachelor’s degrees in those fields (National Center for Education Statistics, 2014). Thus, underrepresentation in the sciences is not for lack of ability or interest (Seymour and Hewitt, 1997; Crisp et al., 2009; Koenig, 2009). Underrepresentation, instead, appears to be a problem of translating the abilities and interests of students from racial/ethnic groups historically underrepresented in the sciences into persistence (Byars-Winston et al., 2011).

The Need for Measures Tested with Groups of Interest
A challenge in understanding factors that impact the persistence of individuals from HU racial/ethnic groups in the sciences is the dearth of theoretically based measures to assess their career and academic development experiences. Although research into the specific factors that account for key academic and career outcomes in the sciences is emerging, many of the published studies have not used measures validated for individuals in the sciences in general or for HU racial/ethnic groups in particular. Moreover, these studies have not reported the psychometric properties of measures used to operationalize their variables or based their measures in theory (Byars-Winston et al., 2011; Trujillo and Tanner, 2014). To identify effective intervention strategies to increase science interests and persistence for individuals from HU racial/ethnic groups, we need measures that are theoretically informed, particularly by vocational and career theory, and psychometrically sound for these populations.

A central element of increasing college students’ interest and persistence in science careers is their participation in undergraduate research experiences (President’s Council of Advisors on Science and Technology, 2012). Extensive data support the positive impact of these experiences on students’ self-reported gains in their research and writing skills, academic self-confidence, research productivity, and increased consideration of graduate degrees in science (see Laursen et al., 2010). However, there is more to know about how such interventions exert their impact on student outcomes. For example, how do programmatic elements in the research training environment impact cognitive factors such as students’ beliefs and identities?

Purpose of the Study
In this study, we validated measures based on social cognitive and science identity theories with a sample of primarily black/African-American and Hispanic/Latino/a undergraduate students who were participating in undergraduate research. We introduced new measures of theoretically derived factors associated with student persistence, including self-efficacy and science identity, and examined potential variability in these factors as a function of gender, racial/ethnic group membership, and the intersection of gender and race/ethnicity. The findings from this study are important for professionals (e.g., instructors, mentors, program directors) working to increase the persistence of HU racial/ethnic groups in science career pathways, because they provide tools to both monitor trainee research-related self-efficacy and science identity and to assess the impact of interventions. The findings will also contribute to future empirical research into the factors in students’ scientific career and academic development that facilitate broadening participation for individuals from HU racial/ethnic groups.

LITERATURE REVIEW
Factors Influencing Persistence in Science
Career paths, including the science career path, are determined by a complex interaction of experiences and choices that begin in early childhood and are continually shaped by personal and environmental factors over the life course. Bakken et al. (2006) argued that current efforts to enhance persistence, particularly in clinical and biomedical sciences, largely focus on environmental elements (e.g., didactic training programs, financial support) and often ignore important person (e.g., race, gender) and cognitive factors (confidence, motivation, identity, etc.) that interact with research training environments. The interplay between a person’s characteristics, his or her cognitions, and the research training environment shapes the person’s career interests, choices, and adjustments (Graham et al., 2013). Two cognitive factors that have been associated with persistence of undergraduates in science are self-efficacy and science identity (Trujillo and Tanner, 2014).

The Self-Efficacy Factor. One of the more well-researched cognitive factors in academic and career development literature is self-efficacy. Self-efficacy (confidence in one’s ability to successfully perform a given task) is a central construct in social cognitive career theory (SCCT; Lent et al., 1994) and is highly correlated with choice of and persistence in a science or engineering major (Lent et al., 1986, 2005).

A model of SCCT that has been modified to incorporate science identity (discussed in the next section) is presented in Figure 1. The gray boxes in the figure indicate the variables measured in the present study. Briefly, SCCT posits that people are likely to form enduring interest in an activity when they view themselves as competent at performing it and when they expect the activity to produce valued outcomes (i.e., when they possess favorable self-efficacy and outcome expectations in relation to the activity). Along with self-efficacy and outcome expectations, personal interests foster particular educational and occupational choice goals (e.g., intentions to pursue a particular career path) that, in turn, make it more likely that people will take actions to achieve their goals (e.g., seek entry into particular academic or occupational paths). Individuals’ subsequent performance attainments (e.g., successes, failures) provide valuable feedback that can strengthen or weaken self-efficacy and outcome expectations and, ultimately, help to revise or stabilize choices.

SCCT proposes that self-efficacy beliefs are formed via four types of information: performance accomplishments (e.g., personal mastery experiences or past successes), vicarious learning (e.g., observing the explicit behaviors of others, such as role models), social persuasion (e.g., verbal encouragement), and affective/emotional arousal experienced while completing a task (e.g., low anxiety, relaxed state). Importantly, the theory takes into account several pathways through which individual...
differences (e.g., gender, race, and personality) affect the academic and career development process.

Bakken et al. (2006) illustrated how SCCT, based on Bandura’s (1997) social cognitive theory, provides a valuable theoretical base for understanding and promoting biomedical and clinical research careers, because it incorporates the reciprocal interactions between person, cognitive, and environmental elements in shaping career outcomes. Studies by Hurtado et al. (2009), Chemers et al. (2011), Estrada et al. (2011), Adedokun et al. (2012), and Byars-Winston et al. (2015) have focused on the contributions of self-efficacy factors to career outcomes for diverse students in the biological and biomedical sciences. Their research has shown support for scientific or research-related self-efficacy (confidence in one’s ability to successfully perform scientific work or conduct research) as an important mediator between students’ research experiences and their subsequent commitment to a research career.

The Science Identity Factor. All individuals have the fundamental need to belong (Allen and Eby, 2007). And that need to belong is often translated into an identification with a given group or context. Identity is defined as the point at which “an individual accepts influence from another person or a group in order to establish or maintain a satisfying self-defining relationship to the other” (Kelman, 2006, p. 3). Research demonstrates that students’ identification (or sense of belonging, fit) with an academic setting is correlated with their academic success and persistence (see Gloria and Robinson Kurpius, 2001). Further, identification with identities that are relevant to a given context, like having a strong identity as an athlete while playing a sport, predicts performance in that context. When a student reports feeling like a scientist, and thereby assumes the identity of scientist, he or she is more likely to pursue a career in science (Estrada et al., 2011). Research shows that identifying as a scientist, or science identity, is related to persistence behaviors (e.g., applying for a PhD degree) (Seymour and Hewitt, 1997; Chemers et al., 2011). Our modified SCCT model in Figure 1 places science identity as a potential intermediary factor, along with relevant self-efficacy and outcome expectancy beliefs, between students’ learning experiences in research and their academic and career outcomes.

Science identity is informed by recognition of one’s self and recognition by others as a potential scientist. Carlone and Johnson (2007) articulated a framework for science identity that includes three overlapping dimensions: competence (knowledge and understanding of science content), performance (skills and opportunities to act like a scientist), and recognition (characterized by acknowledging oneself and being recognized as a scientist by others). Choosing a science career means taking on the identity of a scientist. It is important to note that this science identity must also be negotiated along with other salient social identities (e.g., gender, race, class). As such, the ability to integrate multiple identities plays a role in persistence, too (Hurtado et al., 2009; Laursen et al., 2010; Chang et al., 2011; Chemers et al., 2011; Estrada et al., 2011; Thew et al., 2012).

Studies examining science identity and how it operates to contribute to student academic and career outcomes, particularly for HU racial/ethnic groups, are relatively recent, and there is still more to be learned. For instance, how does a research experience or a research mentor impact a student’s science identity and does science identity moderate those influences on a student’s subsequent science career commitment? Emerging findings from studies by Estrada et al. (2011) and Chemers et al. (2011) suggest that science identity plays a complementary role along with science or research-related self-efficacy beliefs in driving students’ science academic intentions (e.g., intentions to complete a baccalaureate or graduate
degree). Placement of science identity within our modified SCCT model in Figure 1 is based on the findings of Estrada et al. (2011) and Chemers et al. (2011).

**Contribution of This Study to Broadening Science Participation**

There is a published body of work with theoretically informed measures that assess the experiences, attitudes, and beliefs of diverse undergraduate students in science, particularly the work of Chemers et al. (2011) and Estrada et al. (2011). We build and expand upon that important work in several ways. First, whereas the existing science self-efficacy measures assess one’s confidence to perform various research and science-related tasks, we expand our research self-efficacy measure to capture both confidence in performing research tasks and persisting in a research science career pathway (i.e., complete undergraduate degree, pursue and complete graduate degree), following the SCCT work of Byars-Winston et al. (2010) and Lent et al. (2005). Second, although there are validity data for several extant measures of science self-efficacy and science identity, we could not find data on the measurements’ performance (i.e., invariance) with different cultural groups. In this study, we examine validity data regarding our measures’ functioning across diverse groups. Finally, no measures of sources of self-efficacy in a research domain exist. Toward the advancement of full tests of the SCCT model (Lent et al., 1994), such as the one depicted in Figure 1, we developed and report on a new measure of research-related sources of self-efficacy in the present study.

**METHOD**

**Participants**

Since 2001, the American Society for Microbiology (ASM) has sponsored the Annual Biomedical Research Conference for Minority Students (ABRCMS) with continuing funds from the National Institutes of Health (NIH). This conference was designed with the dual purpose of encouraging undergraduate students to pursue advanced training and careers in the biomedical sciences and providing faculty mentors, advisors, and program leaders with resources for facilitating student success. In particular, the specific aims of ABRCMS (www.abrcms.org) are for undergraduate students to

- exchange their research findings and demonstrate scientific expertise;
- prepare for the evolving, global, and interdisciplinary nature of biomedical and behavioral sciences research; and
- transition successfully from undergraduate, postbaccalaureate, and master’s programs to graduate or professional education.

Undergraduate and postbaccalaureate student attendance has doubled since the first year of the conference (2001 = 976; 2013 = 1775), and students consistently represent ~52% of the total conference attendees.

**Data Collection**

Undergraduate and postbaccalaureate students, as well as graduate student and faculty/staff attendees of ABRCMS, are asked to complete a postconference survey on the final day of the conference. The survey is primarily used for formative pur-
TABLE 1. Scale items

**Research self-efficacy**
How much confidence do you have in your ability to:
- Make important contributions to a research team?
- Explain your research topic to other scientists?
- Do research?
- Write a basic research proposal?
- Present a research talk or poster?
- Excel in your science major over the next two semesters?
- Pursue a research science career?
- Complete a science degree?
- Persist with science courses even though you may be a minority in them?
- Pursue a graduate degree in science?
- Complete a graduate degree in science?

**Sources of self-efficacy**
Performance accomplishments
Based on feedback from mentor, in last research experience how well did you:
- Independently conduct experiments or a research project?
- Analyze research data?
- Write a scientific report?
- Prepare a scientific poster or presentation?

Vicarious learning
- My primary research mentor showed me how to conduct a research procedure.
- I look up to my research mentor as a career role model.

Social persuasion
- My research mentor encouraged me to pursue a research science career.
- My research mentor told me I have the ability to be a scientist.

Affective/emotional arousal
- I felt nervous when conducting research.
- I felt anxious about my ability to do research.

**Outcome expectations**
A research science career would allow me to:
- Do work that makes a difference in people’s lives or society.
- Do work that I find satisfying.
- Go into a field with high employment demand.
- Get respect from other people.
- Earn an attractive salary.

**Science identity**
During my most recent research experience, I:
- Felt like a scientist.
- Interacted with scientists from outside of my school.
- Felt part of a scientific community.

recent research experience, with content relating to performance accomplishments (four items), vicarious learning (two items), social persuasion (two items), and affective/emotional arousal (two items). For the performance accomplishments items, participants were instructed to rate how well they had carried out each of the tasks (ranging from 1 = not well at all to 5 = extremely well), and for the remaining domains of items, they rated the degree to which they agreed they had encountered each of the activities (ranging from 1 = strongly disagree to 5 = strongly agree). We acknowledge that context has a large impact on college student experiences and can be variable across both time and across student groups, with greater variability and inconsistency of supports and opportunities for HU students (e.g., Carter, 2006). Because of that variability and inconsistency, we chose to focus on participants’ most recent research experience as opposed to the sum of all their research experiences up to the point of the study in order to assess sources of self-efficacy in one particular context.

**Outcome Expectations.** We used a modified version of the outcome expectations scale from Lent et al. (2005), for which a reliability coefficient of $\alpha = 0.89$ was observed with a sample of racially diverse engineering undergraduate students. We chose five items that captured the three types of outcome expectations defined in SCCT (Lent et al., 1994): physical (e.g., monetary), self-evaluative (e.g., personal satisfaction), and social (e.g., respect from others). Participants were instructed to rate the degree to which they agreed with the various benefits that a research science career would personally provide them, ranging on a scale from 1 = strongly disagree to 5 = strongly agree.

**Science Identity.** The seven items in the science identity scale were drawn from the Attitudes and Behaviors as a Researcher component of the URSSA tool and included practicing authentic scientific inquiry and feeling responsible for a research project (Weston and Laursen, 2015). Weston and Laursen (2015) reported reliability coefficients for this URSSA component of $\alpha = 0.83–0.84$ with a national sample of undergraduate students involved in research. We selected the three items in the component that directly assessed individuals’ interacting with the scientific community and feeling like a scientist. Participants rated their strength of agreement with the items on a five-point scale, ranging from 1 = strongly disagree to 5 = strongly agree. Following our rationale and decision for the context of the sources of self-efficacy scale, participants were instructed to focus on their most recent mentored research experience when responding to the science identity items.

**Data Analysis for Validation of Scales**
The first set of analyses sought to validate the scales with a sample of undergraduate students in the sciences consisting of mainly students from HU racial/ethnic groups. There are four stages of scale development (Netemeyer et al., 2003). The first is the construct definition through theoretical and literature review, the second is the generation of the measurement items, the third is the refinement of scales through testing, and the last is the finalizing of the scales. Stages 1 and 2 are described above, while stages 3 and 4 are covered in the psychometric analyses presented in this paper. Specifically, these analyses explored the internal consistency, and construct validity of each of the four scales (research self-efficacy, sources of self-efficacy, outcome expectations, and science identity).

Scale construct validity and internal consistency were examined through a series of confirmatory factor analyses (CFAs) and use of coefficient alpha for the entire sample of undergraduates. Next, we examined the construct validity as well as the internal consistency of each scale on the individual groups (black/African-American females, black/African-American males, Hispanic/Latino males, and Hispanic/Latina females) to
TABLE 2. Results of CFAs for full sample

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Alpha</th>
<th>RMSEA</th>
<th>SRMR</th>
<th>CFI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research self-efficacy</td>
<td>4.25</td>
<td>0.660</td>
<td>0.910</td>
<td>0.173</td>
<td>0.099</td>
<td>0.831</td>
</tr>
<tr>
<td>Sources of self-efficacy</td>
<td>4.14</td>
<td>0.841</td>
<td>0.673</td>
<td>0.061</td>
<td>0.044</td>
<td>0.960</td>
</tr>
<tr>
<td>Outcome expectations</td>
<td>4.32</td>
<td>0.570</td>
<td>0.788</td>
<td>0.079</td>
<td>0.028</td>
<td>0.986</td>
</tr>
<tr>
<td>Science identity</td>
<td>4.28</td>
<td>0.703</td>
<td>0.730</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>


ensure that the factor structure and internal consistency were acceptable across these diverse groups representing race/ethnicity and gender interactions.

For the initial psychometric analyses, the entire sample of survey respondents was included (N = 688). Of the 688 participants included in this study, there was almost a 2:1 ratio of females (n = 458; 67%) to males (n = 230; 33%). Participants self-reported their racial/ethnic groups as follows: black/African American (n = 288, 42%), Asian American (n = 53, 8%), Hispanic/Latino(a) (n = 275, 40%), Native American (n = 7, 1%), and Pacific Islander/Alaskan Native (n = 5, 0.7%). A smaller number of participants identified as Caucasian (n = 28, 4%) and other (n = 32, 4%). A subsample, described in the following paragraph, was used for the individual group analyses.

Data Analyses for Testing Group Differences

Potential group differences by gender and race/ethnicity were examined as was the intersection of the two statuses. We performed t tests and analysis of variance (ANOVA) to test for differences between the groups. Results and details of these analyses are described below. Large sample sizes allowed a second set of analyses examining group differences on mean scale scores to be performed. A subsample of black/African-American and Hispanic/Latino(a) respondents (N = 563) was evaluated to explore potential racial/ethnic and gender differences in the mean response scores on the scales. The subsample included 197 black/African-American females, 91 black/African-American males, 175 Hispanic/Latina females, and 100 Hispanic/Latino males.

RESULTS

Psychometric Analyses Validate the Scales on a Sample of Undergraduates from Predominantly HU Racial/Ethnic Groups

In this section, we present the results of the analyses exploring the construct validity and the internal consistency of each of the four scales (research self-efficacy, sources of self-efficacy, outcome expectations, and science identity). Results are presented first for the full sample and second for the individual demographic groups.

CFAs Verify Scale Validity

A series of CFAs were run to verify scale validity. These analyses were completed using a maximum-likelihood estimation procedure in the EQS statistical program. We used a combination of measurement properties to estimate and assess model fit using comparative fit statistic (CFI), the standardized root-mean-square residual (SRMR), and the root-mean-square error of approximation (RMSEA; McDonald, 1989; Hu and Bentler, 1999; Jackson et al., 2009). Models demonstrating an RMSEA greater than 0.10 indicate a poor fit, whereas models with an SRMR of less than 0.08 and a CFI value greater than 0.90 indicate a strong fit (Hu and Bentler, 1999). It was not possible to run a CFA with the science identity scale due to the low number of items (three), which resulted in a model with 0 degrees of freedom that did not permit fit statistics to be calculated. An exploratory factor analysis was conducted as an alternate approach to look at the factor structure of the science identity scale.

Full Sample

Results of the CFA for the full sample along with descriptive statistics appear in Table 2. The research self-efficacy model demonstrated an adequate fit, with RMSEA = 0.173, SRMR = 0.099, and CFI = 0.831. There is evidence of a strong fit for the sources of self-efficacy scale, with RMSEA = 0.061, SRMR = 0.044, and CFI = 0.960. The model for the outcome expectations scale demonstrated a strong fit, RMSEA = 0.079, SRMR = 0.028, and CFI = 0.986 (Jöreskog and Sörbom, 1982; Browne and Cudeck, 1993; Hu and Bentler, 1999). We also ran an exploratory factor analysis on the science identity scale for each of the four groups at the intersection of race/ethnicity and gender in our study. Results indicate that the scale items all load on one factor for each of the groups, thus demonstrating that the science identity factor structure is similar across the four groups.

Individual Demographic Groups

We next examined the factor structure of the research self-efficacy, sources of self-efficacy, and outcome expectations scales among four different demographic groups: black/African-American males, black/African-American females, Hispanic/Latino males, and Hispanic/Latina females. Table 3 provides the results for the CFAs conducted on each of the demographic groups.

Research Self-Efficacy. The results show that the research self-efficacy scale has a weak or adequate fit for each of the individual groups examined in this study. For black/African-American males, the CFA models for the research self-efficacy scale demonstrated poor fit (RMSEA = 0.222, SRMR = 0.082, and CFI = 0.772). This was also the case for the CFA model including black/African-American females (RMSEA = 0.203, SRMR = 0.098, and CFI = 0.799). The CFA model for the research self-efficacy scale that included Hispanic/Latino males demonstrated poor fit (RMSEA = 0.146, SRMR = 0.103, and CFI = 0.853), as did the model including Hispanic/Latina females (RMSEA = 0.142, SRMR = 0.117, and CFI = 0.868).

Sources of Self-Efficacy. CFA models examining the sources of self-efficacy scale demonstrated overall good fit for the
individual groups, with the exception of the model that included black/African-American males. For black/African-American males, the CFA model for the sources of self-efficacy scale demonstrated adequate fit (RMSEA = 0.106, SRMR = 0.100, and CFI = 0.906). For black/African-American females, the CFA model for the sources of self-efficacy scale was a strong fit (RMSEA = 0.050, SRMR = 0.048, and CFI = 0.974). The CFA model for the sources of self-efficacy scale that included Hispanic/Latino males had an RMSEA = 0.082, SRMR = 0.060, and CFI = 0.937, all of which indicate a strong fit. This was also the case for the CFA model including Hispanic/Latina females (RMSEA = 0.086, SRMR = 0.067, and CFI = 0.924).

**Outcome Expectations.** Model fit statistics for the outcome expectations scale indicate that the scale has a good fit across each of the four groups we examined, though the model was a better overall fit for black/African-American males and Hispanic/Latino males than for the complementary female samples used in this study. For black/African-American males the CFA models for the outcome expectations scale demonstrated good fit (RMSEA = 0.041, SRMR = 0.029, and CFI = 0.996). The CFA model for the outcome expectations scale also demonstrated good fit for black/African-American females in our sample (RMSEA = 0.134, SRMR = 0.054, and CFI = 0.968). For Hispanic/Latino males, the CFA for the outcome expectations scale demonstrated a strong fit (RMSEA = 0.052, SRMR = 0.037, and CFI = 0.992). The model examining the factor structure of the outcome expectations scale with Hispanic/Latina females had a good fit (RMSEA = 0.115, SRMR = 0.044, and CFI = 0.966).

To summarize, our analyses of each scale via CFA revealed evidence of construct validity among this group of predominantly HU students. Evidence of construct validity was also found for each of the scales for our four demographic groups, with the exception of the research self-efficacy scale.

**Coefficient Alpha Tests Internal Consistency**

For the last step of the psychometric analyses, a series of internal consistency estimates were run using the coefficient alpha (α), which is widely used to establish levels of internal consistency (Netemeyer et al., 2003). This statistic is representative of the consistency (i.e., relationship) between items in a given scale. Values range from 0 to 1, with anything 0.70 and greater being generally considered as good and showing internal consistency, and values between 0.60 and 0.70 being acceptable for parsimonious measurement instruments and scale-development purposes (Netemeyer et al., 2003).

**Full Sample.** Internal consistency statistics for the full sample for each of the four scales are presented in Table 2. The alpha for the research self-efficacy scale was 0.910, demonstrating very strong reliability. The alpha for the sources of self-efficacy scale was 0.673, which is bordering on acceptable. However, this perhaps is not surprising, given that four different dimensions (types of sources of self-efficacy) are being measured with this scale. The alphas for the outcome expectations scale and the science identity scale were 0.788 and 0.730, respectively, both indicating good internal consistency.

**Individual Demographic Groups.** Internal consistency statistics for the research self-efficacy, sources of self-efficacy,
outcome expectations, and science identity scales are reported in Table 3. The alpha coefficient for research self-efficacy is good across all groups, with the exception of Hispanic/Latina females, for which it is fairly low ($\alpha = 0.627$). The alpha coefficients for the sources of self-efficacy scale are good across all groups, though the scale demonstrates the highest level of internal consistency among Hispanic/Latina females. Findings show that the alpha coefficient for outcome expectations is good across all four groups, with the highest level of internal consistency found among black/African-American male participants. Internal consistency statistics for the science identity scale were good across all four groups ($\alpha_{black/African-American male} = 0.783$, $\alpha_{black/African-American females} = 0.701$, $\alpha_{Hispanic/Latino males} = 0.687$, $\alpha_{Hispanic/Latina females} = 0.719$).

**Tests of Group Differences**

The CFA results demonstrated that the factor structure of the sources of self-efficacy and outcome expectations scales hold up within the whole group and within the different populations. The research self-efficacy scale, however, only demonstrated an adequate fit across the whole group and a poor to adequate fit across the individual demographic groups. We also found that the science identity scale is valid for these groups through the use of an exploratory factor analysis. A consistent factor structure means that we have the same variables across different groups. Despite this consistency, there can still be differences between the groups. To determine these differences across different racial/ethnic groups and genders, we performed a series of $t$ tests (gender differences) and ANOVA tests (racial/ethnic differences).

The sources of self-efficacy scale was divided into four sub-scales to evaluate how each of the different groups responded to the sources. The results of these analyses are reported in Table 4. There were no significant differences in mean response levels between males and females for any of the scales. However, findings show that there are significant differences in the mean response levels between black/African-American and Hispanic/Latino(a) respondents on the affective/emotional arousal scale and the science identity scale. Further inspection of the intersection of race/ethnicity and gender revealed statistically significant differences, with a higher mean response on the affective/emotional arousal scale for black/African-American males and a higher mean response on the science identity scale for Hispanic/Latina females compared with the remaining groups.

**DISCUSSION**

Our goal in this study was to use SCCT and science identity theory to validate measures assessing persistence-related factors with students from HU racial/ethnic groups in science majors engaged in research experiences. Students from HU groups are the focus of much discussion regarding broadening, and by “broadening” we mean racially and ethnically diversifying, the talent pool of emerging scientists. We contend that an important step in the science of broadening participation of HU groups in the sciences is the creation and use of valid measures that will allow for greater confidence in the results of investigations into persistence factors for these groups. In considering the contributions of this study, we organize our discussion into persistence factors for these groups. In considering the contributions of this study, we organize our discussion into persistence factors for these groups.
around three primary points: measurement matters, group differences, and implications of the results for practitioners in designing and evaluating programmatic interventions.

Performance of Measures
We have provided a new set of measures to assess HU racial/ethnic students' beliefs about their personal research-related self-efficacy, outcome expectations, and science identities. Analyses of these measures yielded acceptable validity and reliability statistics. Although other studies have used measures of self-efficacy and, to a lesser degree, outcome expectations, we developed the first scale to assess research-related sources of efficacy. This scale now allows for future study of learning experiences that give rise to research self-efficacy beliefs and the learning experiences that give rise to outcome expectations following Fouad and Guillen's (2006) call for more studies of predictive factors for outcome expectations.

The science identity scale exhibited modest but acceptable reliability when tested on our primarily HU racial/ethnic group. While these items were used as a scale cluster in previous studies, they have not been validated on a group of HU undergraduate students in science disciplines. There is room for improvement on this scale. New items could be tested and added to more accurately capture "science identity" for HU racial/ethnic students and subsequently improve reliability. Further validity and reliability studies of the scale with other samples of HU racial/ethnic students and students who are not from HU backgrounds are needed. Such investigations would help to clarify whether the science identity scale items are more or less reliable for all groups of students or whether there is something particular about how the items operate for HU racial/ethnic groups of students.

Interestingly, we found that two sources of efficacy (social persuasion and vicarious learning) loaded onto one factor. Lent et al. (1994) theorized social persuasion (e.g., verbal encouragement) and vicarious learning (e.g., observing the explicit behaviors of others, such as role models) to be two unique sources contributing to an individual's self-efficacy beliefs. However, as our findings demonstrated, the questions intending to measure these two separate sources captured the same underlying factor. Thus, for the individuals in our study, these questions were not distinguishable in their minds, but rather these two sources seem to be conflated into one. That is, observing the behaviors of role models and receiving verbal encouragement operate together or act as a combined source of learning for these individuals. This makes sense, given that mentored research experiences are saturated with observational learning (Laursen et al., 2010), and our data suggest that such learning is potentiated by socially persuasive and encouraging comments. It could also mean that, in the experiences of these individuals, verbal encouragement is only salient when coming from role models they have been observing. By extension, our findings suggest that, for this sample, learning research by observing others model research tasks co-occurs with the verbal encouragement students receive from those they observe. Research mentors would do well to be cognizant that they offer positive encouragement while supporting students' acquisition and mastery of research skills and tasks.

Overall, the results of our validity and reliability analyses indicate that these scales are appropriate for measuring research-related sources of self-efficacy, outcome expectations, and science identity with each of the four demographic groups we examined. Our scales can facilitate future tests of the expanded SCCT model we proposed in Figure 1, which incorporates science identity, examining how the model variables work in relationship with desired student outcomes for HU racial/ethnic students in the sciences. The results from future examinations of the variables can inform mentor interventions, such as identifying what factors build mentees' research self-efficacy and then training mentors to address those factors in their research-mentoring relationships.

Relative to the research self-efficacy scale, the internal reliability coefficients on the total sample and the four demographic groups indicate that items in this scale are consistently measuring the same construct. This is useful information, as we purported to assess both students' perceptions of their ability to perform research tasks and persist in a research science career pathway. However, our examination of measurement invariance, or how the measure performed across groups, revealed that the scale better captures the content of research self-efficacy beliefs of the black/African-American males and females and the Hispanic/Latino males in our study, less so for Hispanic/Latina females. The observed difference for Hispanic/Latina females on the research self-efficacy scale both indicates that more item development and testing are warranted to increase the scale's construct validity and illustrates the importance of future studies examining group differences in their analyses of measures. Our findings also demonstrate that there is variation in the way that each different demographic group experiences the variables assessed in the measures. By extension, our findings also suggest that the influence of the mentored research experience as measured by these four scales may not be experienced in the same way by individuals from these four demographic groups, a point discussed further in the next section.

Group Differences
While we validated the scales in this study using a group that was composed of all racial/ethnic groups (though predominantly HU groups), we took advantage of the size of the data set, which allowed for the examination of both between-group differences, testing the main effects for race/ethnicity and gender, and within-group differences, examining intersection effects of race/ethnicity with gender. In evaluating mean responses to items for black/African-American males and females and Hispanic/Latino/a males and females, we found that black/African-American males reported significantly higher negative affective/emotional arousal for doing research than the other three groups, and Hispanic/Latina females reported significantly higher science identity.

There may be multiple reasons why these two groups emerged with significant mean differences from their counterparts. It may be that the same research training environment is either experienced differently by black/African-American males and Hispanic/Latina females or that the research training environment itself provides different experiences to various groups. Black/African-American men are the only group among HU racial/ethnic groups in STEM fields not making progress in degree attainment (Bidwell, 2015). Their percentage of science and engineering bachelor's degrees has remained essentially
the same in the past decade, from 6.1% in 2002 to 6.2% in 2012 (NSF, 2013). Lack of critical mass, lack of “reliability” (Bidwell, 2015), and being a “minority among minorities” can lead to concerns over potentially fulfilling stereotypes about one’s group (i.e., stereotype threat), all of which can increase one’s anxiety about one’s performance. Hispanic/Latina females, conversely, are not only part of the fastest-growing ethnic groups in the United States but receive more baccalaureate degrees in science and engineering than Hispanic/Latino males (56 vs. 44%; NSF, 2013). Some studies have found that Hispanic/Latina females who are raised in a predominantly patriarchal household are more likely to consider science careers, perhaps owing to fathers’ encouragement of their daughters to consider “male” career options (Bowman, 1993; Marín, 1993). Whether or not Hispanic/Latina females begin their research experiences with higher science identity or elements of the research training environment (e.g., mentors, peers) support and nurture their science identity to a greater degree than it does for other students is unclear.

Overall, our results highlight the gendered nature of racial experiences in science (Lundy-Wagner, 2013). The emergence of group variation in mean responses to the study’s scales suggests that our measures are sensitive enough to detect subtle differences between racial/ethnic and gender groups, specifically for science identity and the sources of self-efficacy. The group differences should be interpreted with caution. Although the results were statistically significant, the present analyses do not provide any data regarding the consequence or impact of these group differences on other variables or outcomes for students. Further investigations can determine whether these differences are replicated with other samples of HU racial/ethnic students in the sciences. Additionally, different analytical methods, including multivariate analyses, can examine whether the group differences are also present in how the variables correlate with each other. For example, how do the research-related sources of self-efficacy differentially explain the variance in research-related self-efficacy and outcome expectancy beliefs? Do these sources of self-efficacy also significantly contribute to science identity?

Implications for Practitioners
The scales we developed and report on herein are important tools for practitioners, such as research program directors and research-based laboratory instructors, who aim to broaden participation in research and the scientific workforce. Practitioners need assessment and evaluation tools that generate valid data about the impact and outcomes of their programmatic and classroom interventions in order to determine whether they are effective for the targeted populations of students. In particular, the nuanced differences in responses between race/ethnicity and gender groups revealed by the analyses presented in this paper reflect important differences in experiences that must be considered when designing and revising interventions.

Beyond the scarcity of evaluation tools available for use by practitioners that have been validated with students from HU racial/ethnic groups, many practitioners, in particular those not trained in the social sciences, are unable to easily select appropriate assessment and evaluation tools. A first step to address this, taken here, is to publish scale development and validation research in journals that reach practitioners and that provide guidance regarding how these tools may be used in the field. Interdisciplinary collaborations between life and social scientists are needed to advance our common goal of broadening participation in science. The research presented here reflects that type of collaboration.

Limitations
The item content captures self-reported ratings of individuals’ beliefs about their abilities, experiences, and identities. They were not designed to be used as formal assessments of students’ abilities and experiences. Whereas the scales for research self-efficacy, outcome expectations, and science identity are worded such that they can be administered as pre or post measures, the content and wording for the sources of efficacy relate to experiences encountered during a specific research training event and, thus, are suited for posttraining measures.

CONCLUSION
There is an increased need for scientific approaches to identifying, implementing, and evaluating initiatives that can advance the academic and career attainment of HU groups in the sciences (Valantine and Collins, 2015). Scientific inquiry into a phenomenon requires quantitative as well as qualitative tools that are valid for the populations of interest experiencing the phenomenon so that there is confidence in the findings. The measures in this study provide tools for investigating the research learning experiences of HU racial/ethnic students and how those experiences contribute to their persistence in science, and information for practitioners to consider about group differences when designing programs.

ACKNOWLEDGMENTS
We gratefully acknowledge the comments and support from Anne-Barrie Hunter, Robert W. Lent, and Nancy Thayer-Hart and help from ABRCMS staff, Clifford W. Houston, and Amy L. Chang. This study was supported with funding from the NIH under grant R01 GM094573 and funding from the Department of Medicine at the University of Wisconsin–Madison. This is independent research, and the views expressed here do not indicate endorsement by the sponsors.

REFERENCES
Measures for HU Science Undergraduates


Early Engagement in Course-Based Research Increases Graduation Rates and Completion of Science, Engineering, and Mathematics Degrees

Stacia E. Rodenbusch,† Paul R. Hernandez,‡ Sarah L. Simmons,§ and Erin L. Dolan†*

†Texas Institute for Discovery Education in Science, College of Natural Sciences, University of Texas at Austin, Austin, TX 78712; ‡Department of Learning Sciences and Human Development, College of Education and Human Services, West Virginia University, Morgantown, WV 26506; §Howard Hughes Medical Institute, Chevy Chase, MD 20815-6789

Submitted March 9, 2016; Revised April 21, 2016; Accepted April 24, 2016
Monitoring Editor: Jennifer Knight

National efforts to transform undergraduate biology education call for research experiences to be an integral component of learning for all students. Course-based undergraduate research experiences, or CUREs, have been championed for engaging students in research at a scale that is not possible through apprenticeships in faculty research laboratories. Yet there are few if any studies that examine the long-term effects of participating in CUREs on desired student outcomes, such as graduating from college and completing a science, technology, engineering, and mathematics (STEM) major. One CURE program, the Freshman Research Initiative (FRI), has engaged thousands of first-year undergraduates over the past decade. Using propensity score–matching to control for student-level differences, we tested the effect of participating in FRI on students’ probability of graduating with a STEM degree, probability of graduating within 6 years, and grade point average (GPA) at graduation. Students who completed all three semesters of FRI were significantly more likely than their non-FRI peers to earn a STEM degree and graduate within 6 years. FRI had no significant effect on students’ GPAs at graduation. The effects were similar for diverse students. These results provide the most robust and best-controlled evidence to date to support calls for early involvement of undergraduates in research.

INTRODUCTION

Undergraduate research experiences (UREs) are seen as integral to training the next generation of scientists, driving governmental and philanthropic agencies to invest millions of dollars annually to support undergraduate research internships (Sadler et al., 2010; American Association for the Advancement of Science [AAAS], 2011; President’s Council of Advisors on Science and Technology [PCAST], 2012). A growing body of research documents the positive outcomes of UREs. Undergraduates who conduct research in science, technology, engineering, or math (STEM) report cognitive gains such as learning to “think and work like a scientist,” affective gains such as finding research enjoyable and exciting, and behavioral outcomes such as increased intentions to pursue further education or careers in science (Seymour et al., 2004; Laursen et al., 2010; Lopatto and Tobias, 2010). An increasing number of well-controlled, large-scale, and longitudinal studies indicate that UREs can attract, retain, and improve the success of undergraduates in STEM (Estrada et al., 2011; Eagan et al., 2013; Hernandez et al., 2013). These results have been the impetus for calls for widespread involvement of undergraduate students in research (AAAS, 2011).

The apprenticeship structure of UREs, in which an undergraduate works one-on-one with a more experienced researcher, such as a faculty member, postdoctoral scientist, or graduate student, limits the number of undergraduates who can participate in research. This limitation, coupled
with interest in expanding the availability and accessibility
of research experiences and the high cost associated with
apprenticeships, has driven the development of courses
that engage students in doing research, also called discov-
ery-based research courses or course-based undergradu-
ate research experiences (CUREs; Wei and Woodin, 2011;
PCAST, 2012; Auchincloss et al., 2014; National Academies
of Sciences, Engineering, and Medicine, 2015). CUREs involve
students in addressing a research question or problem that
is of interest to the scientific community in the context of a
class (Auchincloss et al., 2014). When compared with tradi-
tional lab courses, CUREs afford opportunities for students
to make discoveries that are relevant to stakeholders outside
the classroom, including practicing scientists, and to engage
in iterative work such as troubleshooting, problem solving,
and building off one another’s progress in a way that more
closely resembles the practice of STEM (Auchincloss et al.,
2014; Corwin et al., 2015b).

One example of a national-level, upper-division, single-se-
semester CURE is the Genomics Education Partnership, in
which students enrolled in a genomics-related course finish
raw Drosophila genome sequence data and annotate genes
and other genome features as part of addressing a larger
research question related to Drosophila genome evolution
(Buonaccorsi et al., 2010). The Science Edu-
cation Alliance–Phage Hunters program is an example of a
national-level, introductory, two-semester CURE in which
students identify and characterize novel soil bacteriophages
in the context of a two-semester introductory biology course
series (Halfull et al., 2006; Jordan et al., 2014). Other CURE
models involve addressing a range of research questions
using a common, centrally supported technology, such as
high-throughput sequencing (Buonaccorsi et al., 2011, 2014),
and local CUREs, in which faculty members integrate an as-
pect of their research into courses they teach at their own col-
leges or universities (Bascom-Slack et al., 2012; Klozer et al.,
2013; Harvey et al., 2014).

CUREs have the potential to make research experi-
ences available at scale, rather than to a select few who
seek out research internships or are handpicked by faculty
(Auchincloss et al., 2014). Because CUREs can be offered at
the introductory level, they have greater potential to change
students’ educational and career trajectories than research
internships, which are mostly available to students later in
their undergraduate careers, in junior or senior year. This
enormous potential has led to rapid growth in the number of
CUREs and recommendations for their widespread adop-
tion (AAAS, 2011; PCAST, 2012), despite critiques that point
out the dearth of evidence of their effectiveness and impact
(Linn et al., 2015). Most studies of CURE effectiveness or im-
 pact rely on student self-report of knowledge and skill gains
or intentions to pursue graduate education in STEM or sci-
ence-research related careers, rather than more direct mea-
sures of achievement and retention in STEM. However, sev-
eral CUREs have been in operation long enough to examine
longer-term effects for students—especially whether CURE
participation influences students’ persistence and success in
STEM and in college in general.

The Freshman Research Initiative (FRI) at the University
of Texas at Austin (UT Austin) is a CURE program that was es-

134 CBE—Life Sciences Education

In Course 1, students learn to search and read scientific
literature, and they design and execute one or more sci-
tific investigations, called inquiries, which they summarize
in written and oral reports. During this semester, they also
participate in a matching process through which they are
assigned to a stream. In Course 2, students learn about the
overarching research goals for their stream, complete in-
structional modules to learn concepts and skills specific to
the research, and begin to contribute to the stream’s research.
In Course 3, students become more independent, often pro-
posing and carrying out their own independent subproject
using the skills and understanding they developed in Course
2. Depending on the research, students may either work side
by side on parallel projects or as a member of a team on a
component of the research. As an example, after completing
Course 1, students might join the Supramolecular Sensors
Stream, and make use of spectroscopy, chromatography, or-
ganic synthesis, and biochemical techniques to create and
utilize peptide-based sensors to differentiate wine varietals.
These students can choose to earn either a general biology
or general chemistry lab credit for Course 2 and either inde-
pendent biology or chemistry research credit for Course 3.
Courses 1 and 3 are writing intensive; students who complete
these courses also complete a university writing requirement.

Each section of Course 1 enrolls 25 students and is taught
by a PhD-level lecturer. Each stream (Courses 2 and 3) en-
rolls up to 40 students. These courses are led by a PhD-level
research educator (RE), who is a hybrid of an instructor and
a research scientist hired as a non–tenure-track faculty mem-
ber or postdoctoral associate, and an individual or team of
tenure-track or tenured principal investigators (PIs). A small
number of streams enroll only 15 students per semester and
are led by graduate students who serve in the role of RE.
The RE role is unique and essential to FRI, because each
RE mentors a team of up to 40 undergraduate researchers,
which would not be practical in a more traditional research

The Freshman Research Initiative (FRI) at the University
of Texas at Austin (UT Austin) is a CURE program that was es-

stablished to improve the learning experiences of undergrad-
uates in the College of Natural Sciences (CNS), about half of

whom are life science majors. The program is described in
greater detail elsewhere (Beckham et al., 2015) and summa-
rized here as context for this study. The full FRI program is
a three-course series, which we refer to here as Courses 1, 2,
and 3 for simplicity. FRI students first complete a research
methods course (Course 1), followed by up to two semes-
ters of course-based research (CUREs) in one of 25+ different
areas, called “research streams” (Courses 2 and 3). Current
research streams are offered in a range of science disciplines,
including biology, biochemistry, bioinformatics, chemistry,
computer science, physics, and astronomy (see https://
cns.utexas.edu/fri for a complete list). Students earn three
credit hours for each course, which translates to roughly 9 h
of lab-related work per week. In addition, each course helps
students make progress toward completing their degrees:
Course 1 counts toward university requirements, Course 2
counts as an introductory lab credit, and Course 3 counts as
an upper-division lab or research credit.
FRI was launched with 40 students in 2005 and now serves ~900 students per year, which is ~40% of the incoming class in the CNS. A sufficient number of students have participated in FRI to examine its effectiveness in terms of direct, long-term student outcomes. Specifically, this analysis assessed the degree to which participation in FRI influenced students’ probability of graduating with a STEM degree, probability of graduating within 6 years regardless of major, and educational performance in terms of cumulative grade point average (GPA) at graduation when compared with a matched sample of their peers.

**METHODS**

**Participants**

A sample of 4898 students was drawn from the population of students enrolling at UT Austin between 2006 and 2013 (N = 75,767). This study was designed to test the intermediate- and long-term impacts of the FRI on academic performance and persistence in a STEM major. This study primarily compared students who completed all three semesters of the FRI program with a group of propensity score–matched control students. In this paper, we report data from students first year, junior year, and graduation year (typically fourth or fifth year of enrollment at UT Austin). We restricted the sample to students with complete information for the variables used in the propensity score analysis (N = 53,603; see FRI Program Variables). Students enrolled in programs that guaranteed FRI enrollment were also omitted (i.e., Biology Scholars program, Emerging Scholars program, Women in Science program, Dean’s Scholars Honors program, and Public Health Honors program; N = 52,619). A propensity score–matching procedure was conducted on the resulting sample of FRI (n = 2648) and non-FRI students (n = 49,971).

Finally, the analytical sample used in data analysis was restricted to propensity score–matched FRI and non-FRI students (N = 4898; nFRI = 2449 and nNon-FRI = 2449). About 93% of FRI students had a close propensity score–matched non-FRI student and were thus included in the final analytical sample.

**FRI Program Variables**

The following variables measured “participation” in the FRI program. FRI is a three-course CURE program. Participation in each of the three courses was measured by enrollment data collected from the registrar’s office after the add/drop period ended on the 12th class day of the semester. Participation in Course 1, which students complete in the Fall of their freshman year, was dummy coded (0 = matched control group, 1 = FRI group) for all analyses. Courses 2 and 3 represent the lower- and upper-division research courses of FRI, which students complete in the Spring of their freshman year and Fall of their sophomore year, respectively. Participation in each semester was measured by enrollment data collected from the registrar’s office after the add/drop period ended on the 12th class day of the semester. Spring participation (Course 2) and Fall participation (Course 3) were each dummy coded (Course 2: 0 = did not participate, 1 = participated; Course 3: 0 = did not participate, 1 = participated) for all analyses.

**Identification of Matched Samples of FRI and Non-FRI Participants**

To conduct an analysis of the effect of FRI participation, we first had to identify an appropriate control group of nonparticipating students. We used a propensity score–matching procedure to calculate the probability that a student would be in FRI based on a set of observed covariates in order to correct for selection bias when creating a matched control group (West et al., 2008). The propensity score model (i.e., logistic regression) included 13 variables used in the FRI admissions process to generate a propensity score (from 0 to 1) for each student in the MatchIT software program (Ho et al., 2007, 2011; Thoemmes, 2011). Regarding the variables that influence admissions into FRI, the minimum requirement for entry is a passing score (70%) on a math competency test. Students in several specialty programs in the CNS, such as the Women in Natural Sciences program, are automatically admitted to FRI. These account for ~30% of the FRI population. Students from groups underrepresented in the sciences, such as those with family income less than $40,000 per year, those who are first in their families to go to college, women majoring in physical sciences, computer science, or math, and students with low SAT scores, are also selected for admission. These students account for ~40% of the FRI population. The remaining ~30% of FRI students apply to the program. Applicants are given priority based on their membership in one of the underrepresented groups described above. Finally, there is some attrition from FRI after each semester. Seats that become available in Courses 2 and 3 are filled with students from the applicant waiting list.

We used the following sociodemographic characteristics as matching variables, because they are associated with admission into FRI and persistence in STEM: gender, race/ethnicity, parental education levels, parental income level, and Pell grant eligibility (Schneider et al., 1997; Riegle-Crumb et al., 2012; Supplemental Table S1). We also included variables that have been shown to be associated with enrollment in FRI and students’ choice to major in STEM: SAT total score or ACT equivalent as a measure of prior academic achievement, number of high school science credits earned as a measure of science preparation, and number of high school math credits earned as a measure of math preparation (Wang, 2013). We included the following additional variables in the matching procedure, because they affected students’ likelihood of enrolling in FRI and thus may have resulted in a selection bias: whether students graduated from a Texas or out-of-state high school, the first year students enrolled at UT Austin (e.g., 2006), the first semester students enrolled at UT Austin (entry in Fall is on cycle with FRI admissions), the first college students entered at UT Austin (CNS students are prioritized), and enrollment in the Texas Interdisciplinary Program, a community-building program in the college.

We used FRI students’ propensity scores to identify comparable non-FRI control students (see Supplemental Material for details). The propensity score–matching procedure resulted in two groups of equal size (FRI group n = 2449 and matched control group n = 2449). The percent bias reduction on the matching covariates was 98% in the matched sample (Supplemental Figure S1 and Supplemental Table S2). The following analysis was restricted to matched pairs in which...
the FRI student participated in Course 1 alone \( (n = 416) \), both Courses 1 and 2 \( (n = 882) \), or the complete FRI program \( (i.e., Courses 1, 2, and 3; n = 1151) \), and the non-FRI student participated in no FRI courses. In addition, analysis was restricted to matched pairs in which both students had scores on the outcome and complete data on all predictors.

**Outcomes**

The following variables measured outcomes relevant to participation in FRI.

**Earned Baccalaureate Degree in STEM.** Students who had graduated earned degrees in a variety of colleges \( (e.g., natural sciences, engineering) \). The college of earned degree variable was recoded into a STEM degree dummy-coded variable \( (0 = \text{non-STEM college}, 1 = \text{STEM college}) \), with only the colleges of natural sciences and engineering coded as STEM colleges. Mathematics and computer science degrees are earned from the CNS.

**Earned Any Degree within 6 years of Entry.** Student graduation from UT Austin within 6 years of entry was measured by coding graduation versus nongraduation by Spring of 2015. This variable was dummy coded to represent graduation or nongraduation \( (0 = \text{had not graduated within 6 years of entry}, 1 = \text{graduated with a degree within 6 years of entry}) \). Because our focus was on students who had the opportunity to graduate within 6 years, this analysis was restricted to students in our data set entering UT Austin on or before 2009 \( (i.e., we had graduation data for students up to Spring 2015) \).

**Cumulative GPA at Graduation.** Cumulative college GPA was measured at graduation. Cumulative GPA was measured on a scale from 0 to 4.

**Cumulative GPA at Midpoint of College Tenure.** Cumulative college GPA was measured at the midpoint of the undergraduate college tenure \( (i.e., \text{Fall of junior year}) \). Cumulative GPA was measured on a scale from 0 to 4.

**Control Variables.** All covariates used in the propensity score-matching process were also used as variables in the regression analyses to control for chance imbalances across groups \( (\text{Schafer and Kang, 2008}) \). Control variables included: gender \( (\text{female, male}) \), race/ethnicity \( (\text{Asian, Hispanic, white, or other}) \), enrollment in Texas Interdisciplinary Program \( (\text{yes, no}) \), SAT total score \( (\text{or ACT equivalent}) \), Pell grant eligibility \( (\text{yes, no}) \), number of units of science on high school transcript, number of units of math on high school transcript, how students were initially accepted into UT Austin \( (\text{transcript, number of units of math on high school transcript, eligibility \( (\text{yes, no}) \), number of units of science on high school transcript, SAT total score \( (\text{or ACT equivalent}) \), Pell grant}) \), how students were initially accepted into UT Austin \( (\text{transcript, number of units of math on high school transcript, eligibility \( (\text{yes, no}) \), number of units of science on high school transcript, SAT total score \( (\text{or ACT equivalent}) \), Pell grant}) \), whether students were in the complete program \( (\text{yes, no}) \), maternal and paternal education levels \( (\text{less than college degree, college degree \( [2 \text{ or 4 years}], \text{or advanced degree}) \), parental income level \( (\leq \text{39,999}, \text{40,000 to 79,999}, \text{80,000 to 99,999}, \text{or \geq 100,000 per year}) \), and first college entered at UT Austin \( (\text{STEM-related college = natural sciences or engineering, or non-STEM-related college such as education or business}) \).

**Treatment of Missing Data**

For the sample of 4898 participants, 12.7% of participants were missing data on their midcollege cumulative GPA, 43.0% were missing data for their cumulative GPA at graduation \( (i.e., \text{had not yet graduated by Summer 2015}) \), 41.6% were missing data on their major at graduation and time to degree completion \( (i.e., \text{had not yet graduated by Summer 2015}) \). Our matched sample included both those who had time to graduate \( (i.e., 4 \text{ years for traditional students or} \text{2 years for transfer students}) \) and a smaller number of those who did not \( (i.e., \text{their first year enrolled was 2012 or 2013}) \). Although those who did not have time to graduate \( (\text{and their matched control}) \) did not contribute to analyses related to any of the graduation outcomes \( (i.e., \text{cumulative GPA, STEM degree, 6-year graduation rate}) \), they were retained because they contributed to the analysis of the FRI effect on midpoint GPA, which was a suspected mediator of the FRI effect on cumulative GPA \( (\text{see Supplemental Material for details}) \).

To ensure unbiased estimates of the effect of FRI, we only used whole linked pairs of participants in which both the FRI and matched control participant provided data for the analysis. This approach restricted our analytical sample of the STEM degree and cumulative GPA outcomes to cases in which both members of the matched pair \( (i.e., \text{both the FRI student and the matched counterpart}) \) graduated in or before Summer 2015 \( (\text{regardless of the number of years to degree}) \). This approach also restricted our analytical sample of the 6-year graduation outcome to cases in which both the FRI student and the matched counterpart started at UT Austin on or before 2009 and thus had the opportunity to graduate within 6 years \( (e.g., \text{for those starting in Fall 2006, graduation by Summer 2013; for those starting in Fall 2009, graduation by Summer 2015}) \). To account for missing data and to account for chance imbalances on covariates used to estimate the propensity scores, we controlled for all covariates used in the propensity score matching in our regression models of the FRI treatment effect \( (\text{Enders, 2010}; \text{Pan and Bai, 2015}) \).

Finally, it is important to note that, even with missingness, all of our analyses were more than adequately powered to detect small effects. An a priori power analysis indicated that the sample size required to detect a small effect \( (i.e., \text{odds ratio} = 1.50) \) of FRI on STEM degree and 6-year graduation was \( N = 776 \), while the sample size required to detect a small effect \( (i.e., R^2 = 0.02) \) on cumulative GPA was \( N = 476 \) \( (\text{Faul et al., 2007}; \text{Chen et al., 2010}) \).

**RESULTS**

**Graduation with STEM Degree**

We assessed students’ attainment of a STEM degree based on descriptive statistics \( (\text{Table 1}) \) and bivariate correlations \( (\text{Table 2}) \) and found a raw difference favoring the FRI group. However, raw differences between FRI and non-FRI groups may be untrustworthy, as they do not control for chance imbalances on the matching covariates and they use data from unlinked members of matched pairs \( (e.g., \text{one member of the matched pair graduated with a STEM degree \( [\text{STEM degree} = 1] \), but the other member of the matched pair had not yet graduated \( [\text{STEM degree} = \text{missing}]) \). Therefore, we conducted a logistic regression analysis on all matched pairs with graduation data \( (\text{both pairs graduated; STEM degree} = 0 = \text{non-STEM college}; 1 = \text{STEM-related college}) \) to determine the effect of FRI participation on students’ probability of graduating with a STEM degree. We used a hierarchical approach in the logistic regression analysis \( (\not\text{not}) \)
Research Courses Increase Persistence

To make these findings more concrete, we calculated the predicted probability of earning a STEM degree for students in the non-FRI control and FRI groups. After controlling for other factors in the model, non-FRI students had a 71% predicted probability of graduating with a STEM degree compared with 94% for FRI students who completed all three courses (47% of FRI students completed all three courses; Figure 1A). Students who only participated in Course 1 (17% of FRI students completed only Course 1) or Courses 1 and 2 (36% of FRI students completed Courses 1 and 2) were just as likely to graduate with a STEM degree as non-FRI students (O.R. Course 1 = 0.69; O.R. Courses 12 = 1.37).

Graduation within 6 years

Descriptive statistics and bivariate correlations indicated a slight raw difference in students’ 6-year graduation rate favoring the FRI group (Tables 1 and 2). To assess an unbiased effect of FRI participation on students’ probability of graduating within 6 years of entering college regardless of major, we conducted a logistic regression analysis on matched pairs in which both had the opportunity to graduate within 6 years: FRI students and matched controls who both enrolled at UT Austin on or before 2009. We used the same parameter estimates in the final step of the logistic regression model revealed that students who participated in all three semesters of FRI (Courses 1, 2, and 3) were significantly more likely to graduate with a STEM degree compared with the non-FRI control group (O.R. Courses 123 = 6.08, 98.3% CI [3.66, 10.12]; see Supplemental Table S3 for complete details). To make these findings more concrete, we calculated the predicted probability of earning a STEM degree for students in the non-FRI control and FRI groups. After controlling for other factors in the model, non-FRI students had a 71% predicted probability of graduating with a STEM degree compared with 94% for FRI students who completed all three courses (47% of FRI students completed all three courses; Figure 1A). Students who only participated in Course 1 (17% of FRI students completed only Course 1) or Courses 1 and 2 (36% of FRI students completed Courses 1 and 2) were just as likely to graduate with a STEM degree as non-FRI students (O.R. Course 1 = 0.69; O.R. Courses 12 = 1.37).

Table 1. Summary of descriptive statistics of outcomes and key predictors as a function of FRI status

<table>
<thead>
<tr>
<th>Variable</th>
<th>FRI</th>
<th>Matched control</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>%</td>
</tr>
<tr>
<td>STEM degree</td>
<td>1482</td>
<td>81</td>
</tr>
<tr>
<td>6-year graduation</td>
<td>1082</td>
<td>79</td>
</tr>
<tr>
<td>Cumulative GPA</td>
<td>1482</td>
<td>81</td>
</tr>
<tr>
<td>FRI course work</td>
<td>2449</td>
<td>100</td>
</tr>
<tr>
<td>Course 1 only</td>
<td>416</td>
<td>17</td>
</tr>
<tr>
<td>Courses 1 and 2</td>
<td>882</td>
<td>36</td>
</tr>
<tr>
<td>Courses 1, 2, and 3</td>
<td>1151</td>
<td>47</td>
</tr>
<tr>
<td>Midpoint GPA</td>
<td>2188</td>
<td>81</td>
</tr>
</tbody>
</table>

STEM degree codes: 0 = non-STEM, 1 = STEM; 6-year graduation codes: 0 = did not graduate, 1 = graduated within 6 years. Course variables (e.g., Course 1 only) were dummy coded to indicate level of participation in FRI = 1 versus otherwise = 0 (reference group was the non-FRI matched control group).

Table 2. Summary of bivariate correlations among outcomes and key predictors

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 STEM degree</td>
<td>1</td>
<td>0.13**</td>
<td>0.08**</td>
<td>-0.09**</td>
<td>0.01</td>
<td>0.21**</td>
<td>0.10**</td>
</tr>
<tr>
<td>2 6-year graduation</td>
<td>1</td>
<td>0.09**</td>
<td>0.09**</td>
<td>-0.10**</td>
<td>-0.01</td>
<td>0.14**</td>
<td>0.45**</td>
</tr>
<tr>
<td>3 Cumulative GPA</td>
<td>1</td>
<td>0.09**</td>
<td>-0.09**</td>
<td>-0.01</td>
<td>0.12**</td>
<td>0.98**</td>
<td></td>
</tr>
<tr>
<td>4 Course 1 only</td>
<td>1</td>
<td>0.14**</td>
<td>-0.14**</td>
<td>-0.17**</td>
<td>-0.11**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Courses 1 and 2</td>
<td>1</td>
<td>-0.26**</td>
<td>-0.01</td>
<td>1</td>
<td>0.15**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 Courses 1, 2, and 3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 Midpoint GPA</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

STEM degree codes: 0 = non-STEM, 1 = STEM; 6-year graduation codes: 0 = did not graduate, 1 = graduated within 6 years. Course 1 codes: 0 = matched control, 1 = FRI; Courses 2 and 3 codes: 0 = did not participate, 1 = participated.

**p < 0.01.
The results indicated that completing the full FRI program has a statistically significant effect on students’ probability of graduating within 6 years, over and above control variables (Table 3). Parameter estimates in the final step of the logistic regression model revealed that students who participated in all three semesters of FRI were significantly more likely to graduate within 6 years (O.R. Courses123 = 2.43, 95% CI [1.34, 4.43]; Supplemental Table S4). To make these findings more concrete, we calculated the predicted probability of graduating within 6 years for students in the non-FRI control and FRI groups. After controlling for other factors in the model, non-FRI students had a 66% predicted probability of graduating with any degree within 6 years compared with 83% for FRI students (Figure 1B). FRI students who only participated in Course 1 or Courses 1 and 2 were just as likely to graduate within 6 years as non-FRI students (O.R. Course1 = 0.63; O.R. Courses12 = 1.07).

Cumulative GPA

Again, descriptive statistics and bivariate correlations indicated a slight raw difference in cumulative graduation GPA, favoring the FRI group (Tables 1 and 2). To assess an unbiased effect of FRI participation on educational performance at graduation, we conducted a regression analysis on all matched pairs with cumulative graduation GPA scores. Preliminary analysis indicated an FRI effect on midpoint GPA (Supplemental Table S5). Thus, midpoint GPA was entered in step 3 as a potential mediator of the effect of participating in FRI. As above, the results indicated that FRI membership (step 2) had a statistically significant effect on cumulative GPA at graduation, over and above control variables (Table 3). FRI students who completed Courses 1 and 2 or all three courses exhibited statistically significantly higher graduation GPA compared with the non-FRI control group (step 2; \( b_{\text{Courses12}} = 0.07 \) and \( b_{\text{Courses123}} = 0.12 \), but students who completed only FRI Course 1 (step 2; \( b_{\text{Course1}} = 0.01 \)) were not significantly different from the non-FRI control group. We suspected that grades in FRI courses themselves could be influencing cumulative graduation GPA. Thus, we controlled for midpoint GPA and found that the positive effects of participating in FRI were nullified (Figure 1C and Supplemental Table S6).

Potential Race, Gender, and First-Generation Moderation Effects

We explored whether students from different backgrounds differed in their outcomes as a result of participating in FRI.

---

Table 3. Regression analysis (logistics or OLS) with graduation with a STEM degree, graduation within 6 years with any degree, and graduation cumulative GPA as outcomes

<table>
<thead>
<tr>
<th>Step</th>
<th>Predicator</th>
<th>Graduation with STEM degree (n = 1624)</th>
<th>Graduation within 6 years (n = 990)</th>
<th>Graduation cumulative GPA (n = 1510)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>–2LL</td>
<td>Pseudo-R²</td>
<td>( \Delta \chi^2(df) )</td>
</tr>
<tr>
<td>1.</td>
<td>Controls*</td>
<td>1653.38</td>
<td>0.183</td>
<td>217.79 (24)**</td>
</tr>
<tr>
<td>2.</td>
<td>FRI courses</td>
<td>1545.68</td>
<td>0.266</td>
<td>107.70 (3)**</td>
</tr>
<tr>
<td>3.</td>
<td>Midpoint GPA</td>
<td>0.951</td>
<td>0.723</td>
<td>21,705.91 (1)**</td>
</tr>
</tbody>
</table>

\(-2LL = -2\times\log \text{likelihood}; \text{pseudo-}R^2 = \text{Nagelkerke } R^2 \text{ estimate of effect size}; \Delta \chi^2 = \text{change in overall chi-square from the previous model.}

*The list of control variables is described in the Methods section. ***p ≤ 0.001.
Specifically, we tested whether students’ race/ethnicity, gender, or first-generation college status moderated the effect of FRI on the outcomes. Exploratory moderated regression analyses (logistic and OLS) indicated that students’ sociodemographic characteristics did not moderate the effects of FRI on outcomes. Given the analytical sample sizes, number of predictors in our models, and the adjusted alpha level, our exploratory analyses were more than adequately powered to detect small moderating effects (i.e., O.R. = 1.50 or $R^2 = 0.02$; power > 0.99).

**DISCUSSION**

To the best of our knowledge, this is the largest and most carefully controlled analysis to date of the effects of participating in a CUREs on long-term student outcomes that are of high interest to students and institutions alike. Specifically, the data reported here indicate that participation in early CUREs significantly increases students’ likelihood of graduating with a STEM degree and graduating within 6 years. After controlling for other variables, the outcomes of participating in the full FRI program were the same regardless of students’ gender, race/ethnicity, and first-generation college status, showing that these effects were robust for diverse students. Results from these analyses demonstrate the importance of using quasi-experimental techniques for controlling for selection bias in determining the effects of research experiences, since the data show that the variables that influenced entry into FRI had statistically significant effects on all of the outcomes we examined.

The effects of FRI differed depending on whether students completed Courses 1, 2, and 3, which could be due to the nature of the courses or to time spent in the program. In Course 1, students have total freedom to define their own investigations, from posing questions to investigate to designing studies to collecting and analyzing data to constructing and evaluating scientific arguments. Courses 2 and 3 are more similar to UREs, because students engage in conducting novel studies that build on and contribute to a faculty member’s ongoing research, with the potential to yield publishable results as well as methods, data, and other products (e.g., inventions, companies) that are of interest to communities outside the classroom. Thus, the problem space has been defined to some extent. Students carve out their own aspect of the research to pursue and must collect and analyze data and construct and evaluate arguments but may not have complete latitude to select their research questions or methods. This study provides a preliminary test of whether having full intellectual responsibility posing research questions is important for students to achieve desired outcomes (National Academies of Sciences, Engineering, and Medicine, 2015). The parameter estimates from our regression models (Supplemental Tables S4–S6) indicated that Course 1 alone did not have a significant effect on any of the outcomes we examined, yet model fit was improved by including Course 1 in all three models (Table 3). These results suggest that investigatory courses like Course 1 may have distinct positive effects on graduating with a STEM degree when compared with research courses (i.e., Courses 2 and 3). Alternatively, it may be that the independent effects of each FRI course on students’ probability of graduating with a STEM degree can simply be attributed to longer exposure to a learning environment that is more motivating than traditional lab course experiences (Graham et al., 2013).

The distinct, significant effects of Courses 2 and 3 on students’ likelihood of graduating in 6 years and graduating with a STEM degree indicate that the duration of students’ involvement in CUREs is important for their outcomes. Specifically, the data indicate that a one-semester research course is sufficient to achieve these outcomes to some extent but that participation in additional semesters is important for maximally realizing these outcomes. This finding adds to those from Shaffer and colleagues (2014), who found that students who spent more time on their CURE work reported increased learning and greater interest in STEM courses and in STEM in general. These results are likely to be conservative estimates of the effect of participating in CUREs, because the bivariate correlations show that participation in Courses 1, 2, and 3 are all fairly highly correlated. It is likely that collinearity between participating in each course suppresses the independent effects of each course. Larger samples of students who participate in Course 1 only or Courses 1 and 2 only are needed to confirm this.

These analyses were conducted with data from a CURE program that has involved enough students for a sufficient length of time to examine long-term outcomes such as graduation rates and majors. The extent to which these results will apply to other CUREs, especially CUREs that enroll students later in their undergraduate degrees, needs to be determined by conducting similar, carefully controlled studies. Given that many CUREs are small in scale or have more finite life spans, this may prove difficult. An alternative approach would be for studies of CUREs to report long-term outcomes of participating and nonparticipating students such that meta-analyses can be done in the future to identify effects across research course experiences.

These findings are arguably the most robust evidence to date that CUREs improve the outcomes of undergraduate STEM students. We have statistically controlled for background variables related to academic motivation and preparation (e.g., prior achievement, math and science preparation, parental education) and controlled for initial entry into FRI. This lends confidence that the outcomes reported here can be attributed to CURE participation. However, there are likely to be other variables not included in our analysis that may predict FRI participation and cause the outcomes of interest. We are currently collecting data on psychological variables that may predict students’ participation in FRI and their persistence in college and in STEM (e.g., motivation, interest in research; Hernandez et al., 2013) in order to more fully understand the effects of CURE participation per se.

These results do not yield insights into the features of CUREs that lead to these outcomes. There are many structural differences between FRI and traditional lab courses that could be leading to the outcomes reported here (Auchincloss et al., 2014). For example, Courses 2 and 3 meet in dedicated lab spaces that become a sort of scientific home for students. Typically, two wet-lab FRI groups meet in a single large lab space, such that up to 80 students are cycling in and out of the space over the course of the week. Students working on computational projects meet in regularly scheduled conference-style classrooms or a robotics lab and also work online at a distance. The lab spaces are open to students and staffed.
by REs, graduate or undergraduate teaching assistants, or peer mentors throughout the day. FRI lab spaces often become a place where students not only conduct research but also study for classes and spend more time informally. The involvement of undergraduate mentors gives students access to near peers who have recent experience learning the research and who can provide general advice on navigating the first 2 years of college. Class size is not likely to be a major factor, since enrollments are similar between FRI courses and standard laboratory courses, and most FRI courses enroll up to 35 students, which is larger than the typical 24-person introductory lab course. Different versions of FRI that make use of curricular and instructional staffing models are now being implemented at universities across the country. Cross-site study of student outcomes has the potential to yield insight into which FRI design elements are necessary and sufficient to achieve the results reported here.

Future research on CUREs should focus on using research and theory from social sciences, including situated learning (Brown et al., 1989), communities of practice (Wenger, 1999; Lave and Wenger, 1991), and knowledge integration (Linn et al., 2015), to understand the features of CURE design and implementation that lead to these long-term outcomes (Corwin et al., 2015a). Recent research aimed at distinguishing CUREs from traditional lab courses indicates that the extent to which students have opportunities to make discoveries that are of broad interest, engage in iterative work and sufficiency to achieve the results reported here.

Future research on CUREs should focus on using research and theory from social sciences, including situated learning (Brown et al., 1989), communities of practice (Wenger, 1999; Lave and Wenger, 1991), and knowledge integration (Linn et al., 2015), to understand the features of CURE design and implementation that lead to these long-term outcomes (Corwin et al., 2015a). Recent research aimed at distinguishing CUREs from traditional lab courses indicates that the extent to which students have opportunities to make discoveries that are of broad interest, engage in iterative work and sufficiency to achieve the results reported here.

FRI represents a scalable, affordable way to meet this demand. According to predicted probabilities in this study, out of every 100 students who enter college, 17 more will complete an undergraduate degree if they complete FRI. For every 100 students who graduate, 23 more will stay in a STEM major if they complete FRI. A rough estimate of the total per-student cost of FRI is $500 per student for 8–10 wk of research internships and to the tuition dollars lost when students leave college. Costs could be lowered further by scaling up some of the cost-saving measures that we have implemented at UT Austin, such as offering peer mentors relevant course credit instead of pay, hiring senior undergraduates instead of graduate students as teaching assistants, or hiring graduate students as REs. Other models should also be tested, such as tenure-track or tenured faculty serving as the RE as part of their standard teaching responsibilities.

Given that FRI boosted retention among students regardless of their background, the diversity of students enrolled in the program provides the additional benefit of diversifying to the STEM workforce. In the long term, growing a more diverse STEM workforce has the potential to produce more creative, effective, and feasible ideas than would be accomplished by homogenous groups (McLeod et al., 1996). In the near term, FRI can be a model for addressing the massive attrition of undergraduate students from STEM disciplines and ensuring that all students have the potential to earn higher wages and experience lower unemployment rates associated with STEM-related jobs (U.S. General Accounting Office, 2005; Langdon et al., 2011; PCAST, 2012).
ACKNOWLEDGMENTS
We thank the many PIs and REs who provide research leadership and instruction in FRI, including the FRI faculty (Eric Anslyn, Dean Aplling, Karen Browning, Andrew Ellington, Ronny Hadani, Kristen Harris, Christine Hawkess, Graeme Henkelman, Bradley Holliday, Vishwanath Iyer, Richard Jones, Thomas Juenger, Alan Lloyd, Jeffrey Luci, John Markert, Stephen Martin, Risto Mikkulainen, Jon Robertus, Stanley Roux, Neal Rutledge, Paul Shapiro, Scott Stevens, Keith Stevenson, Peter Stone, Claus Wilke, and Don Winget) and the FRI Research Educators and Research Methods Instructors (Joshua Beckham, Jared Bowden, Brandon Campetilla, Grace Choy, Gregory Clark, Art Covert, Anson D’Aloisio, Lauren DePue, Vivian Feng, Eman Ghanem, Antonio Gonzales, Bradley Hall, A. Katie Hansen, Gregory Hatlestad, Richard Heineman, Todd Hether, Kathryn Kavanagh, Patrick Killion, Joel Lehman, Matteo Leonetti, Marsha Lewis, Albert MacKrell, Michael Montgomery, Gregory Palmer, Jeremy Paster, Mary Poteet, Kristen Procko, Michael Quinlan, Stuart Reichter, Timothy Riedel, Moriah Sandy, Mithra Sathishkumar, G. Christopher Shank, Ruth Shear, Gwendolyn Stovall, Samuel Taylor, Daniel Tennant, Anne Tibbetts, Alona Varshul, Travis White, and Liang Zhang). We also thank Jane Huk for assistance with data collection and cleaning and Sarah Eddy, Scott Freeman, Catherine Riegle-Crumb, and Christopher Runyon for technical help and feedback on the manuscript. This work was supported by the CNS, a National Science Foundation award (NSF CHE 0629136), and two Howard Hughes Medical Institute (HHMI) grants (52005907 and 52006958). The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of NSF or HHMI. This study was reviewed and determined to be exempt by the Institutional Review Board at UT Austin (protocol 2014-11-0086). Sarah L. Simmons was the inaugural director of the Freshman Research Initiative and was employed by UT Austin when the data for this study were collected. Funding for FRI from the HHMI preceded Simmons’ employment there.

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


Riegle-Crumb C, King B, Grodsky E, Muller C (2012). The more things change, the more they stay the same? Prior achievement fails to explain gender inequality in entry into STEM college majors over time. Am Educ Res J 49, 1048–1073.


Colleagues as Change Agents: How Department Networks and Opinion Leaders Influence Teaching at a Single Research University

T. C. Andrews, ** E. P. Conaway, † J. Zhao, ‡ and E. L. Dolan§

†Department of Genetics and ‡ Department of Sociology, University of Georgia, Athens, GA 30602; §Texas Institute for Discovery Education in Science, University of Texas, Austin, TX 78712

Submitted August 6, 2015; Revised January 4, 2016; Accepted January 19, 2016

Monitoring Editor: Elizabeth Ambos

Relationships with colleagues have the potential to be a source of support for faculty to make meaningful change in how they teach, but the impact of these relationships is poorly understood. We used a mixed-methods approach to investigate the characteristics of faculty who provide colleagues with teaching resources and facilitate change in teaching, how faculty influence one another. Our exploratory investigation was informed by social network theory and research on the impact of opinion leaders within organizations. We used surveys and interviews to examine collegial interactions about undergraduate teaching in life sciences departments at one research university. Each department included discipline-based education researchers (DBERs). Quantitative and qualitative analyses indicate that DBERs promote changes in teaching to a greater degree than other departmental colleagues. The influence of DBERs derives, at least partly, from a perception that they have unique professional expertise in education. DBERs facilitated change through coteaching, offering ready and approachable access to education research, and providing teaching training and mentoring. Faculty who had participated in a team-based teaching professional development program were also credited with providing more support for teaching than nonparticipants. Further research will be necessary to determine whether these results generalize beyond the studied institution.

INTRODUCTION

College biology instructors are being asked to reconsider traditional teaching strategies in favor of evidence-based teaching strategies (Freeman et al., 2014). The most prominent call for reform in the life sciences, called Vision and Change, urges faculty to focus on teaching core concepts and competencies and to integrate more opportunities for students to be active participants in their learning and in the practice of science (American Association for the Advancement of Science, 2011). Instructors have also been encouraged to teach in ways that increase student motivation and metacognition, which in turn can influence students’ learning and achievement in science (e.g., Glynn et al., 2007; National Research Council [NRC], 2012; Tanner, 2012; Stanton et al., 2015). Despite these calls for changing undergraduate biology instruction, traditional teaching strategies remain common (NRC, 2012), and little is known about how to help college biology faculty reform their teaching to align with recommendations.

Colleague-colleague relationships have the potential to facilitate reform in undergraduate biology education. Colleagues are embedded in the same institutional context, so they can help one another navigate local constraints to achieve change (Henderson et al., 2011). Collegial relationships are also enduring and thus are well positioned to provide the long-term support needed to change one’s teaching (e.g., Andrews and Lemons, 2015). Colleagues are accessible to faculty who may not seek teaching support outside their departments or institutions. Furthermore, interpersonal relationships—especially face-to-face interactions—are important
in catalyzing changes in strongly held beliefs (Rogers, 2003), which is needed to engender meaningful and sustained change in teaching (Henderson et al., 2011).

Existing teaching reform efforts already take advantage of collegial relationships. For example, the National Academies Summer Institutes on Undergraduate Education, a 5-d workshop on evidence-based teaching strategies, enrolls faculty only in departmental teams in order to build collegial support networks that continue when participants return to their institutions. Faculty learning communities (FLCs), which are small groups of faculty who meet regularly over an extended period of time with the objective of improving their teaching (Cox, 2004), are also designed to take advantage of collegial support. FLC participants have reported that the experience helps them change their teaching, for instance, by building their confidence in experimenting with different teaching strategies (Tovar et al., 2015) and promoting use of student-centered teaching practices (Polich, 2008; Light et al., 2009).

Academic colleagues provide other forms of support for one another’s work, but support for teaching may be especially important. Colleagues provide mentoring, ties to powerful people, emotional support, and opportunities for professional advancement (e.g., van Eck Peluchette and Jeanquart, 2000; van Emmerik and Sanders, 2004). Support provided by colleagues is particularly influential for pretenure faculty (e.g., van Emmerik and Sanders, 2004; Pifer and Baker, 2013), but faculty at later career stages also benefit from interactions with colleagues (van Eck Peluchette and Jeanquart, 2000). Collegial support for teaching may be even more important than support for other aspects of faculty work, such as research, because graduate and postdoctoral training in science often fail to provide adequate preparation for teaching (Austin, 2002; Richlin and Essington, 2004; Foote, 2010; Hunt et al., 2012). Faculty end up learning about teaching “on the job,” where colleagues may be the primary sources of information, educational materials, inspiration, and encouragement.

Theoretical traditions from other disciplines provide insight regarding the role of social relationships in education reform. The construct of “opinion leaders” is a potentially useful lens for investigating how colleagues impact one another’s teaching practices. This construct hails from communication studies, a discipline with a rich history of investigating how individuals within organizations adopt innovations (e.g., Rogers, 2003). From this perspective, evidence-based teaching strategies can be considered “innovations” and academic departments are “organizations.” Most organizations have opinion leaders who are particularly influential in whether or not other members of an organization adopt an innovation (Rogers, 2003). Opinion leaders are able to influence others’ attitudes or overt behavior through informal interactions, and they do so with relative frequency (Rogers, 2003). In an organization oriented to change, opinion leaders tend to be innovators, whereas organizations opposed to change may have opinion leaders who are strong advocates of traditional practices (Rogers, 2003). We posit that there are departmental opinion leaders who have significant potential to influence their colleagues’ views about teaching and teaching practices.

Social network theory allows for the examination not only of who the key players are within an organization (e.g., opinion leaders) but also of which organizational relationship are important for sharing resources and producing outcomes (Burt, 1992; Granovetter, 1973). We focus in particular on identifying the networks related to undergraduate teaching within departments (i.e., who is interacting with whom about teaching) and characterizing the quality and influence of the relationships that bind each network together (i.e., what resources are shared, and to what end). As Cross and Borgatti (2004) argue, this approach has the potential to reveal individual and departmental “potential to recognize, assimilate, and take action on new problems or opportunities” (p. 138), namely, to improve undergraduate teaching.

Colleague–colleague relationships could be leveraged to encourage and aid faculty in changing their teaching and improving student learning. However, we lack empirical evidence regarding the nature and outcomes of colleague–colleague interactions about undergraduate teaching in the life sciences. Without this knowledge, we are ill equipped to help faculty and departments take advantage of the influence and assistance colleagues could provide to one another, and we are likely missing opportunities to positively impact undergraduates’ learning and success.

Here, we present an exploratory study that examined how life sciences faculty at a research university interact with and benefit from departmental colleagues regarding undergraduate teaching. The department is a promising system for changing teaching in academia, because teaching assignments and course development generally occur within departments, hiring and retention decisions begin in the department, and departments have their own cultures and practices (e.g., Austin, 1996; Silver, 2003; Wieman et al., 2010). Therefore, we conducted an in-depth study of colleague–colleague interactions in four life sciences departments within a single institution. This institution is categorized by the Carnegie classification system as having “very high research” activity, which is the most research-focused designation for higher education institutions (Carnegie Foundation for the Advancement of Teaching, 2011). We began our investigation of the role of colleague–colleague relationships in teaching reform at this type of university because research-focused institutions are expected to have environments particularly unsupportive of teaching (e.g., Rice and Austin, 1990; Knorek, 2012). Other institution types, such as small liberal arts colleges, have a stronger record of dedicating institutional energy and resources toward supporting faculty as teachers (e.g., Rutz et al., 2012). Investigating this single institution allowed us to study more than 110 faculty members in multiple departments while controlling for differences that might exist across institutions.

This study system was also suitable, because each department we studied employed tenure-track discipline-based education researchers (DBERs). Hiring DBERs in science departments is becoming increasingly common (e.g., Bush et al., 2008, 2011), but little is known about the impact of DBERs within departments. Two goals that have been described for these hires are improving undergraduate science education and cultivating departmental culture change toward a focus on education in the sciences (Bush et al., 2015). Understanding how faculty perceive, interact with, and are affected by DBERs will help us better understand the impact of these types of hires. We did not specifically design this study to investigate DBERs in departments but rather to study collegial relationships in departments that include...
DBERs, so we did not explicitly ask participants about interactions with DBERs. Therefore, it is a powerful investigation of whether and how DBERs are perceived to influence faculty to change their teaching.

We used a mixed-methods approach, collecting and analyzing both survey and interview data, to address the following research questions:

1. To what degree do colleague–colleague interactions about undergraduate teaching occur and between whom?
2. Who provides resources for undergraduate teaching as a result of collegial interactions?
3. Who influences colleagues to change their views and behaviors related to undergraduate teaching (i.e., who are opinion leaders), and how have they influenced colleagues to change?

METHODS

Participants

We collected data from faculty in four departments using an online survey in Fall 2013. We also collected data from faculty in three of the four departments through one-on-one interviews in Spring and Summer 2014. These departments employ very few non-tenure-track faculty, so the results presented here are limited to assistant, associate, and full professors, as well as emeritus faculty who were still active in the department at the time of data collection. We invited all faculty in the four target departments to participate by email (n = 113). We launched a friendly competition among departments by promising home-baked treats at three consecutive faculty meetings for the department with the highest response rate. For the interviews, we strategically recruited a sample of professors of all ranks. To maintain the confidentiality of our participants, we have assigned each department a color as a pseudonym. The University of Georgia Institutional Review Board determined that this study met the criteria for exempt review procedures.

Quantitative Data Collection and Analysis

Survey Data Collection. We collected survey data to gain comprehensive, quantitative data from faculty regarding who interacted with whom about undergraduate teaching, what resources they garnered from these interactions, and what outcomes they experienced as a result of these interactions. Using an online survey in Fall 2013, we asked faculty in the four target departments to select their department and then presented a list of departmental colleagues and asked participants to: “Please select the people in your department with whom you interact about undergraduate biology. Interacting includes everything from talking about teaching to sharing resources to receiving and giving feedback.” For subsequent survey questions, each respondent saw only the names selected in this first question. We asked respondents to indicate how often they interacted with each person and “what outcomes you have experienced as a result of interacting with this person about teaching undergraduate biology.” Respondents could indicate as many of the listed items as were applicable (Table 1). Within this list, it is useful to differentiate “resources,” which are something that can be received in a social interaction, and “changes” occurring as a consequence of a social interaction. Respondents could indicate that they “engaged in conversation” with the individual or experienced “no outcomes” as a result of interacting with the individual; these responses are not considered in our analyses as they do not represent resources or change (Table 1).

We also asked respondents to “please list any other people employed at [your institution] with whom you interact about teaching undergraduate biology at least monthly.” Respondents could list up to five people. Respondents reported the resources received and changes experienced as a result of interacting with these colleagues.

Other Data Collection. We collected publicly available data about each faculty member in all four departments to complement survey data. We determined academic rank in Fall 2013 (i.e., assistant, associate, or full professor), gender, and whether or not a faculty member was a DBER. We defined DBERs as faculty whose jobs have a research expectation related to biology education. We also determined whether each faculty member had participated in a regional teaching professional development program using a list provided by the program organizers. This program is relevant to our investigation, because it accepts teams from within a department and may therefore facilitate the development of teaching-related relationships among departmental colleagues. Hereafter, we refer to this program as “Regional Teaching PD.”

Social Network Analysis. We used social network analyses (SNA) to determine the degree to which faculty interacted with their colleagues about teaching (research question 1), including the characteristics of their teaching-related networks and what individual attributes were associated with interactions. SNA is useful for elucidating the social structure of a community, including the determinants of the structure of relationships among members of the community (Hawe et al., 2004; Grunspan et al., 2014). In this study, we treated each department as a community or network. Networks consist of actors, and any two actors within a network are a dyad (Hawe et al., 2004; Grunspan et al., 2014). The relationships between actors are called ties, and the focus of our analysis is on ties. SNA allows us to investigate the characteristics of an individual faculty member that are associated with interacting with relevant respondents to indicate how often they interacted with each person and “what outcomes you have experienced as a result of interacting with this person about teaching undergraduate biology.” Respondents could indicate as many of the listed items as were applicable (Table 1). Within this list, it is useful to differentiate “resources,” which are something that can be received in a social interaction, and “changes” occurring as a consequence of a social interaction. Respondents could indicate that they “engaged in conversation” with the individual or experienced “no outcomes” as a result of interacting with the individual; these responses are not considered in our analyses as they do not represent resources or change (Table 1).

We also asked respondents to “please list any other people employed at [your institution] with whom you interact about teaching undergraduate biology at least monthly.” Respondents could list up to five people. Respondents reported the resources received and changes experienced as a result of interacting with these colleagues.

Other Data Collection. We collected publicly available data about each faculty member in all four departments to complement survey data. We determined academic rank in Fall 2013 (i.e., assistant, associate, or full professor), gender, and whether or not a faculty member was a DBER. We defined DBERs as faculty whose jobs have a research expectation related to biology education. We also determined whether each faculty member had participated in a regional teaching professional development program using a list provided by the program organizers. This program is relevant to our investigation, because it accepts teams from within a department and may therefore facilitate the development of teaching-related relationships among departmental colleagues. Hereafter, we refer to this program as “Regional Teaching PD.”

Social Network Analysis. We used social network analyses (SNA) to determine the degree to which faculty interacted with their colleagues about teaching (research question 1), including the characteristics of their teaching-related networks and what individual attributes were associated with interactions. SNA is useful for elucidating the social structure of a community, including the determinants of the structure of relationships among members of the community (Hawe et al., 2004; Grunspan et al., 2014). In this study, we treated each department as a community or network. Networks consist of actors, and any two actors within a network are a dyad (Hawe et al., 2004; Grunspan et al., 2014). The relationships between actors are called ties, and the focus of our analysis is on ties. SNA allows us to investigate the characteristics of an individual faculty member that are associated with interacting with
colleagues about teaching, and the characteristics of faculty members in a dyad that make an interaction about teaching more likely.

Preparing the Network Data. We constructed a network matrix for each department using responses to the survey question asking respondents to indicate which departmental colleagues they interacted with about teaching. Each matrix included all faculty in the department. Nonresponse can be an issue in SNA, because it can result in missing network information. However, a lack of response does not mean that actors are left out of the network, because respondents can report ties to nonrespondents (Borgatti and Molina, 2003). Missing data on ties can be reconstructed through a process called “symmetrizing” that assumes a tie exists if it is reported by at least one member of a dyad (Stork and Richards, 1992). Before symmetrizing, it is important to determine whether assuming reciprocal ties is reasonable. For our study, the phrasing of our survey question supports symmetrizing, because we asked respondents to report “with whom you interact” about undergraduate teaching, which implies undirected or symmetrical ties (i.e., if I interact with you, then you interact with me). This is distinct from a directed network, such as a mentoring network (e.g., if I mentor you, you do not necessarily mentor me). In addition, determining whether an interaction has occurred is relatively objective. In contrast, asking respondents to nominate “best” friends would rely heavily on individual perceptions and judgments, meaning two participants would be less likely to agree on whether they were best friends or not. Stork and Richards (1992) suggest that assuming reciprocal ties is justified if actors who responded to the survey (respondents) are similar to actors who did not respond to the survey (nonrespondents). We followed their approach for comparing respondents and nonrespondents, and the results support our decision to symmetrize the data (see the Supplemental Material for detailed descriptions of these comparisons).

Even after symmetrizing the data, we were missing data for some dyads, because neither person responded to the survey. We quantified this as “missingness,” or the percent of dyads for which we have no data out of all possible dyads in a network (Table 2). Missing data have a negative effect on network mapping and model estimation, because missing data can lead to underestimating the strength of relationships. According to Costenbader and Valente (2003), the ranking of an actor based on being nominated is relatively stable even at a low sampling level, such as 50% missingness. All four of our networks have missingness below 50%, but we chose to proceed conservatively. The Yellow department had 44% missingness, meaning that we lacked data from both members of a dyad for 44% of the possible dyads within the department. Therefore, we calculated simple network metrics, but did not fit complex models for this department. The other departments were missing data from no more than 30% of the possible dyads in the department (Table 2).

Calculating Simple Metrics to Characterize Networks. We used network density and transitivity to characterize collegial interactions in each department. One of the most widely used concepts in network analysis is “density.” Density describes the general level of linkage among actors in a network, and it is one way to quantify group cohesion. The more actors who are connected to one another, the denser the network will be, and the more cohesive the network will be. Evidence has shown that a dense web of interactions facilitates both information exchange and knowledge transfer (Burt, 1987; Reagans and McEvily, 2003). To measure network density, we calculated the incidence of interactions among all possible dyads in a network (Friedkin, 1981) using the software UCINET (Borgatti et al., 1991). It is defined as the proportion of ties that exist out of the total number of ties possible. The formula for the density is

\[
\text{Density} = \frac{1}{n(n-1)/2}
\]

where \(l\) is the number of ties present and \(n\) is the number of actors the network contains.

This measure can vary from 0 to 1, with no connections among actors in a network corresponding to a density of 0 and every actor being connected to every other actor corresponding to a density of 1.

Another well-known small-scale attribute of social networks is transitivity, which is defined as the tendency of two actors to be connected to each other if they share a mutual neighbor (Holland and Leinhardt, 1971). Transitivity has been nicely described in the conventional wisdom that “friends of my friends are my friends.” Transitivity deals with triads, which are connections among three actors. With undirected network data like ours, there are four possible types of triadic relations (no ties, one tie, two ties, or all three ties). Counts of the relative prevalence of these four types of relations across all possible triads provides a sense of the extent to which a network is characterized by “isolation,” “couples only,” “structural holes” (i.e., where one actor is connected to two others who are not connected to each other), or “clusters.” We are particularly interested in the proportion of triads that are transitive, compared with triads that are two-legged (Figure 1). Two-legged triads could be thought of as “potentially transitive” triads, because they could be made transitive with just one more tie.

Transitivity implies interactions are largely confined to friends and to friends of friends. High transitivity tends to reinforce group norms, and low transitivity allows for multiple viewpoints (Quar nachus and Henderson, 2015). A meaningful way to measure transitivity for each network (i.e., department) is to calculate the number of closed (i.e., transitive) triads divided by the sum of all closed triads and all two-legged triads (i.e., potentially transitive). Transitivity ranges between 0 and 1; it is 1 for a perfectly transitive network, where all triads are closed. Seen in this way, a

<table>
<thead>
<tr>
<th>Department</th>
<th>Reciprocity (%)</th>
<th>Missingness (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green</td>
<td>55</td>
<td>12</td>
</tr>
<tr>
<td>Yellow</td>
<td>54</td>
<td>44</td>
</tr>
<tr>
<td>Red</td>
<td>48</td>
<td>30</td>
</tr>
<tr>
<td>Blue</td>
<td>43</td>
<td>18</td>
</tr>
</tbody>
</table>

*Reciprocity within data from survey respondents was sufficiently high to assume reciprocity among all departmental colleagues. Missingness is calculated after assuming reciprocity and indicates the percent of ties in a network for which we have no data.
transitivity index of 0.4 means 40% of all transitive and two-legged triads are transitive.

Exponential Random Graph Models. We used exponential random graph models (ERGMs) to make statistical inferences about the likelihood of observing particular patterns of interactions as a result of chance. Building on statistical exponential family models (Strauss and Ikeda, 1990), ERGMs recognize the lack of independence among ties between actors in the data. For instance, the relationship between actor A and actor C may be contingent upon the relations between actor A and actor B. So instead of focusing on individuals as the unit of analysis, ERGMs predict the likelihood of an interaction (i.e., a tie being present) based on individual, dyadic, and local network properties. For example, faculty who engaged in the Regional Teaching PD may have more interactions (i.e., ties) with departmental colleagues than would be expected if relationships occurred completely at random. This is an individual-level property of a tie. Dyadic properties take into account how the individual characteristics of both members of a tie are associated with the presence of a tie, such as whether female faculty are more likely to have ties with other female faculty than with male faculty. ERGMs can include parameters representing extraneous structural forces, such as tendencies to reciprocate ties or complete a two-legged triad. These processes contribute to interaction patterns found in the data beyond the individual characteristics and dyad dynamics.

We used the network and ergm packages in R for this analysis. The ergm package allows maximum-likelihood estimates of ERGMs to be calculated using Markov chain Monte Carlo (Robins et al., 2007; Goodreau et al., 2008; Hunter et al., 2008). We fit separate models for the Red, Blue, and Green departments. We assumed that teaching-related ties were generated by local social processes and that these social processes may depend on the surrounding social environment (i.e., on existing relations). To reflect all the social processes that might contribute to the interaction patterns observed in our data, we examined the following individual attributes: gender, academic position, position type (i.e., DBER or non-DBER), and participation in Regional Teaching PD. To examine the same variables at a dyadic level, we included node (i.e., actor) match terms in our models to capture the effect of the two members of a dyad having the same attributes. The phenomenon of people interacting with people similar to them at higher rates than those who are dissimilar is called homophily. For example, we tested whether junior faculty reported interacting at a higher rate with other junior faculty compared with their senior counterparts. We also included an edges term, which controls for the baseline probability of a tie, or the proportion of ties in a network that exist out of all possible ties.

Regression Analysis. All regression analysis was completed in R, version 3.1.2. We used two generalized linear regression models to determine the attributes associated with providing resources for teaching (research question 2), and the attributes associated with prompting colleagues to change their views about teaching and teaching practices (research question 3). We counted each type of resource that a faculty member was reported to provide to each survey respondent. A faculty member who was reported to provide feedback to three survey respondents was counted as providing three resources, as was a faculty member who provided three different resources (i.e., feedback, instructional materials, and social support) to a single survey respondent. We used the same strategy to calculate the number of changes a faculty member was reported to have caused.

Both regression models included a common set of explanatory variables: gender, academic rank, position type, and participation in Regional Teaching PD. The total number of resources a faculty member was reported to provide or the number of change occurrences caused may be impacted by the number of survey respondents from that department, which varied from 10 to 16 across departments. We included department as an explanatory variable in our models to control for this potential influence. We also controlled for whether or not a faculty member was a survey respondent by including this variable. We hypothesized that the exchange of resources would predict changes in teaching or views about teaching. Therefore, the model examining change included the number of resources provided as an additional explanatory variable.

We used existing literature and exploratory analyses to decide what regression analyses were most appropriate for achieving our study objectives, given the nature of our data. We asked faculty to report their interactions within their departments, creating a lack of independence among data collected from survey respondents in the same department. Mixed models can account for this sort of hierarchical data structure by including the nesting factor (i.e., department) as a random effect. However, we studied four departments, and in cases with fewer than five groups, multilevel modeling does not provide an advantage over a classical regression model that includes groups as fixed effects (Gelman and Hill, 2007). Therefore, we controlled for department by including it as an explanatory variable in both models.

Regression Models. We used a zero-inflated Poisson model for the teaching resources data (UCLA: Statistical Consulting Group, n.d.). A zero-inflated negative binomial model includes two different models: a binary model to account for excess zeros and a count model. The binary model improves the model fit but does not provide results that inform our research questions. We include the estimated intercept from this model in our results table (see Table 7 later in the article) but do not discuss it further. We present the back-transformed regression coefficients and confidence intervals. We calculated bootstrapped 95% confidence intervals using the bias-corrected and accelerated (BC) method with 10,000 replicates (Efron, 1987). We used a Poisson model to examine the associations between individual attributes and reports.
of causing change. We back-transformed regression coefficients and calculated 95% confidence intervals.

**Qualitative Data Collection and Analysis**

**Interview Data Collection.** We conducted hour-long, one-on-one interviews with faculty from three of the four departments in order to acquire rich descriptive data from faculty regarding colleague–colleague interactions about undergraduate teaching. The interviews were conducted as part of a larger investigation of the working environments in which college biology faculty make decisions regarding undergraduate teaching. The interviews were semistructured, meaning that the interviewers asked the same questions of all participants while remaining flexible about the order of questioning and prompting interviewees for more detail as needed (full protocol in the Supplemental Material). Although the interview protocol did not specifically ask about interactions with departmental colleagues, many of the questions elicited information about the faculty members whom interviewees perceived to be most influential regarding teaching. Two questions elicited a greater number of relevant quotes than the others: “What is going on in your department that helps you to be a good teacher?” and “Is there anyone in your department who you would say is particularly influential to undergraduate teaching?” Both T.C.A. and E.P.C. conducted the interviews, which were audio-recorded and transcribed verbatim for analysis.

Two of the authors are DBER faculty, and that fact was not hidden from research participants. We took several steps to mitigate the potential bias this could introduce to our work. Aside from our names on recruitment communications and on the consent form, there was nothing in the survey to cue respondents to think about DBER faculty. At no point in the survey or interviews did we ask any questions specifically about DBER faculty (see the Supplemental Material). One of the interviewers (T.C.A.) was a DBER faculty member, and this could have cued interview participants to think about DBER faculty more than they otherwise would have. However, this investigator had been in her position for less than a year when the interviews were conducted, and she did not conduct any interviews with faculty she had previously met. Interviews with faculty with whom she had previously interacted were scheduled and conducted by the other interviewer. We opted not to collect interview data from the department to which she belonged, which is why we present interview data from only three departments.

**Qualitative Data Analysis.** Qualitative analysis allowed us to elucidate how colleagues influence change in teaching views and behaviors (research question 3).

**Identifying Opinion Leaders.** We used the results of the survey data to identify faculty members who were reported to prompt the most change in views about teaching or teaching practice. To identify the faculty who engendered the most change, we calculated the median and SD of the number of change occurrences attributed to each faculty member by survey respondents in each department. We then calculated the number of change occurrences equal to the median plus two times the SD. Any faculty member in a department who was reported to have caused a number of change occurrences equal to or higher than this value was considered an “opinion leader” and was investigated further. In each department, the opinion leader with the smallest number of change occurrences still had at least twice as many change occurrences as the next highest count in the department. For example, in one department the three opinion leaders were credited with causing seven, nine, and 11 change occurrences, whereas the next most influential faculty member was credited with only three change occurrences. This provides further support to our claim that department members perceived these individuals as distinctive in the degree to which they influence teaching. We limited our qualitative investigation to only those faculty who were reported as contributing to substantial change in the department.

**Extracting Relevant Data.** We then used Atlas.ti to analyze portions of interview transcripts in which participants discussed the contributions of the identified opinion leaders. We conducted this analysis for the Yellow, Green, and Red departments only (i.e., those for whom we had collected interview data). We used two strategies to identify sections of interview transcripts that discussed opinion leaders. First, previous coding undertaken for the larger project included identifying and marking any section of text that mentioned a DBER faculty member, which was useful, because many of the opinion leaders were also DBER faculty. The goal of the previous analysis was much broader than simply identifying sections of text about opinion leaders, so our strategy was more thorough than would have been necessary simply to identify quotes relevant to this study. Three researchers worked to code transcripts, and at least two researchers worked on the analysis of each transcript. We first coded as a team, discussing to reach consensus regarding which sections of text referred to DBER faculty. Later a single researcher analyzed a transcript, and then a second researcher read the transcript and checked each code to determine whether codes should be added or removed or whether the section of text designated by a code included all relevant and no extraneous information. If the second researcher disagreed with any part of the analysis conducted by the first researcher, they discussed until they reached consensus. Interviews analyzed in this way were divided so that the researcher who checked the analysis was the person who had conducted the interview. This was useful, because the second researcher was familiar with the interview and therefore was able to efficiently identify ideas that the initial researcher had missed and also to question the interpretation of data.

Our second strategy included searching the text of all transcripts for any place that an opinion leader’s name was mentioned and marking these sections of text. Not every mention of an opinion leader was useful, because some did not focus on social relationships among faculty. For example, there were quotes about the hiring and promotion process for DBER faculty that were not relevant to this study. Rather, we focused on quotes about opinion leaders that were relevant to our research questions.

**Thematic Text Analysis.** With a complete list of quotes about opinion leaders in hand, we conducted thematic text analysis to detect patterns of ideas expressed by interviewees. We sought to compare opinion leaders with one another to identify similarities and differences within this group. We also aimed to identify similarities and differences among departments by comparing opinion leaders in one department with those in
the other two. We analyzed data at two different grain sizes to accomplish these aims. We read and reread full quotes, which can be considered “raw” data, and labeled these quotes to indicate themes. We created a cross-comparison table by generating short synopsis statements to represent each quote. Each row of the table is one case (i.e., opinion leader), and the columns of the table represent major themes in the data set. The identity of these themes emerges and is refined as researchers add synopsis statements to the table. Essentially, the table is an organizational structure for identifying patterns in the data and allowed us to better compare opinion leaders.

All qualitative analyses were iterative and collaborative. One researcher (T.C.A.) spearheaded the analysis process, and another researcher (E.P.C.) cross-checked the data and results repeatedly throughout the process. Cross-checking involved reading all quotes and comparing them with the synopsis statements and the Results section of the manuscript with the goal of identifying anything that was missing, incomplete, or extraneous. T.C.A. and E.P.C. engaged in a number of discussions about the emerging themes throughout this process. Additionally, three experienced qualitative researchers reviewed the analysis and results at multiple points in the process.

Quotes about opinion leaders often included their names, which we have replaced with pseudonyms. To protect the anonymity of our participants, we do not attribute quotes to specific interviewees. We assigned all DBER opinion leaders pseudonyms beginning with the letter “D” and all non-DBER opinion leaders pseudonyms beginning with the letter “N” to allow readers to better recognize the differences between group members were perceived by their colleagues. All of our qualitative analysis focused on opinion leaders, so the results apply to that select group only. Some quotes have been lightly edited for clarity and grammar. We have also excised sections of quotes that are not relevant for the sake of brevity, as well as sections that have the potential to reveal the identity of interviewees or opinion leaders.

RESULTS

This exploratory study examined faculty within one research institution. Response rates to the survey ranged from 33% of faculty in the Yellow department to 63% of faculty in the Green department, with an average survey response of 46% across all four departments (Table 3). These response rates are similar to those achieved in another investigation collecting network data from college faculty (Quardokus and Henderson, 2015). Response rates for interviews were somewhat higher, with an average of 57% of contacted faculty participating in an interview (Table 3). The data presented in this study have been collected from a total of 59 faculty members: 27 provided both survey and interview data, 25 completed only the survey, and seven only participated in an interview.

**Research Question 1: To What Degree Do Colleague–Colleague Interactions about Undergraduate Teaching Occur and between Whom?**

Across departments, survey respondents reported interacting with a median of 18% (SD = 19%) of their departmental colleagues about undergraduate teaching, but this value varied from 14% (SD = 21) in the Blue department to 30% (SD = 22) in the Red department. Interactions about undergraduate teaching did not occur frequently. More than 60% (n = 33) of survey respondents had no departmental colleagues with whom they interacted about undergraduate teaching on a weekly basis, and 31% (n = 16) had no departmental colleagues with whom they interacted about teaching even once per month (Table 4). Survey respondents also reported few out-of-department colleagues with whom they interact about teaching regularly. Most survey respondents (n = 36, 68%) did not list anyone outside their departments with whom they interact at least once per month about undergraduate teaching. Just nine survey respondents (17%) reported interacting with out-of-department colleagues about undergraduate teaching on a weekly basis.

One group that stood out as well connected to out-of-department colleagues was faculty holding DBER positions. In fact, 57% of the reported social interactions with out-of-department colleagues were with DBER faculty from another department. About half of these connections were between a DBER and a non-DBER faculty member, and the other half were between two DBER faculty. These results indicate that DBER faculty are strongly connected to one another and are also responsible for many of the interdepartmental interactions about undergraduate teaching. Two staff from the Center for Teaching and Learning were each nominated once as an out-of-department colleague. Some nominations reflected personal relationships among faculty, such as spouses. Other relationships may have formed as a result of coparticipation in university-wide teaching-focused groups.

Overall, network data suggest that colleague–colleague relationships regarding teaching are relatively common in these departments, at least compared with other science departments. Network density ranged from 0.15 to 0.34. In

<table>
<thead>
<tr>
<th>Department</th>
<th>Faculty contacted for survey</th>
<th>Survey participants</th>
<th>Survey response rate (%)</th>
<th>Faculty contacted for interview</th>
<th>Interview participants</th>
<th>Interview response rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green</td>
<td>19</td>
<td>12</td>
<td>63</td>
<td>19</td>
<td>13</td>
<td>68</td>
</tr>
<tr>
<td>Yellow</td>
<td>43</td>
<td>14</td>
<td>33</td>
<td>27</td>
<td>14</td>
<td>52</td>
</tr>
<tr>
<td>Red</td>
<td>23</td>
<td>10</td>
<td>44</td>
<td>21</td>
<td>11</td>
<td>52</td>
</tr>
<tr>
<td>Blue</td>
<td>28</td>
<td>16</td>
<td>57</td>
<td>None</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Total</td>
<td>113</td>
<td>52</td>
<td>46</td>
<td>67</td>
<td>38</td>
<td>57</td>
</tr>
</tbody>
</table>

*Faculty were recruited from four life sciences departments for an online survey and from three departments for one-on-one interviews.
comparison, a teaching network study of five science depart- ments at another research-intensive university found network densities ranging from 0.06 to 0.216 (Quardokus and Henderson, 2015). It is important to note that density can be affected by both the size of the network (i.e., as networks grow in size, density declines) and response rate (i.e., missing ties may contribute to a less dense network). Even taking this into account, there were differences among departments (Table 5). For example, the Red and Blue departments have a similar number of faculty (Red: 23; Blue: 28) and the response rate was higher for Blue (57%) than for Red (44%) (Table 3). Yet the density for Red department was nearly double that of Blue department (Table 4), indicating that the group cohesion about undergraduate teaching is greater for Red than for Blue.

Transitivity was highest in the Yellow and Red departments (Table 5). Recall that transitivity for a network is measured by the number of transitive triads divided by the sum of transitive triads and two-legged triads (Figure 1). The department with the lowest transitivity (Blue) had a transitivity score similar to departments investigated by Quardokus and Henderson (2015). The Green and Red department are similarly sized and have a similar density of ties, but transitivity is much higher in the Red department.

Network analysis using ERGMs revealed that both individual and dyadic attributes predicted the likelihood of a tie between departmental colleagues in some departments.

**Table 4. Frequency distribution of interactions with departmental colleagues**

<table>
<thead>
<tr>
<th>Number of colleagues interacted with</th>
<th>Monthly or more often (%)</th>
<th>Weekly or more often (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>16 (31)</td>
<td>33 (63)</td>
</tr>
<tr>
<td>1</td>
<td>12 (23)</td>
<td>10 (19)</td>
</tr>
<tr>
<td>2</td>
<td>9 (17)</td>
<td>4 (8)</td>
</tr>
<tr>
<td>3</td>
<td>2 (4)</td>
<td>3 (6)</td>
</tr>
<tr>
<td>4</td>
<td>5 (10)</td>
<td>1 (2)</td>
</tr>
<tr>
<td>5</td>
<td>3 (6)</td>
<td>1 (2)</td>
</tr>
<tr>
<td>6+</td>
<td>5 (10)</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

*Count (percent) of survey respondents (*n* = 52) who report interacting with departmental colleagues at least once per month and at least once per week.

The odds of male faculty interacting with colleagues was 0.35 times (95% CI: 0.18–0.67; *p* < 0.001) and 0.34 times (95% CI: 0.16–0.70; *p* < 0.01) lower than the odds of female faculty interacting with colleagues about undergraduate teaching in the Red and Green departments, respectively. Similarly, we observed that, in the same two departments, faculty who had participated in Regional Teaching PD were 1.97 times (95% CI: 1.14–3.42; *p* < 0.05) and 3.03 times (95% CI: 0.83–11.06; *p* < 0.09) more likely to interact with colleagues about undergraduate teaching than those who did not participate in this training. In only one department (Green) was academic rank associated with the likelihood of interactions regarding undergraduate teaching. Full and associate professors were 2.27 times (95% CI: 1.06–4.88; *p* < 0.001) and 7.3 times (95% CI: 2.56–21.29; *p* < 0.001) more likely to interact with colleagues about undergraduate teaching than were assistant professors, respectively. None of these attributes was associated with the likelihood of interacting with colleagues about teaching in the Blue department. Additionally, we did not find that having a DBER position increased the likelihood of interacting with colleagues about undergraduate teaching in any department. One potential explanation for this null finding is that there were only one or two DBER faculty per network (department), making the odds of identifying significant effects on this variable low.

Our models also estimated the likelihood of collegial interactions based on dyadic characteristics, such as gender homophily (individuals tend to interact more with same gender others than the opposite gender). We examined all the characteristics (gender, academic rank, participation in Regional Teaching PD, position type) at the dyadic level, but only one variable revealed a pattern. Rank homophily was present among associate professors in the Blue department; associate professors were 6.82 times (95% CI: 1.26–36.8; *p* < 0.05) more likely to interact with other associate professors in their departments than with colleagues of a different rank. In contrast, Green department full professors demonstrated a tendency against rank homophily. The likelihood of full professors interacting with one another was only about one tenth of the likelihood of them interacting with faculty of other ranks (95% CI: 0.02–0.52; *p* < 0.05).

**Research Question 2: Who Provides Resources for Undergraduate Teaching as a Result of Collegial Interactions?**

We asked survey respondents to indicate whether they received the following resources as a result of interacting with colleagues about undergraduate teaching: instructional materials, social support, feedback, and information (Table 1). Social support and information were the resources most commonly reported (Table 6). Respondents reported receiving no resources from 24% of the colleagues with whom they interacted about undergraduate teaching. The median number of resources that faculty were reported to have provided within their departments was 4 (SD = 6), and this value ranged from 0 (*n* = 31) to more than 20 (*n* = 4).

After we controlled for department, several characteristics were significantly associated with the number of resources provided to colleagues: position type (i.e., DBER...
Survey participants (n = 52) selected all outcomes they had experienced as a result of interactions with each departmental colleague with whom they reported interacting about undergraduate teaching.

vs. non-DBER), academic rank, and having participated in this study (Table 7). Holding other variables constant, DBER faculty were reported to provide 2.09 times more resources for colleagues than non-DBER faculty. Compared with assistant professors, associate professors were reported to have provided 2.64 times more resources and full professors were reported to have provided 1.70 times more resources. Survey respondents were reported to provide 1.90 times as many teaching resources for faculty than nonrespondents (Table 7). We hypothesize this association is attributable to an unmeasured construct, such as interest in teaching, which makes a faculty member more likely to respond to a survey about undergraduate teaching and more likely to engage in productive interactions about teaching. The number of resources a faculty member was reported to provide did not vary by gender, department, or participation in Regional Teaching PD.

### Table 6. Frequency of outcomes reported to result from collegial interactions, ordered most to least common

<table>
<thead>
<tr>
<th>Resources</th>
<th>Percent of outcomes reported</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social support</td>
<td>25</td>
</tr>
<tr>
<td>Information</td>
<td>24</td>
</tr>
<tr>
<td>Useful feedback</td>
<td>18</td>
</tr>
<tr>
<td>Instructional materials</td>
<td>14</td>
</tr>
<tr>
<td>Changes</td>
<td></td>
</tr>
<tr>
<td>Changed teaching practices</td>
<td>11</td>
</tr>
<tr>
<td>Changed views about teaching</td>
<td>9</td>
</tr>
</tbody>
</table>

Research Question 3: Who Influences Colleagues to Change Their Views and Behaviors Related to Undergraduate Teaching (i.e., Who Are Opinion Leaders), and How Have They Influenced Colleagues to Change?

Ultimately, we are interested in colleagues as catalysts for improving undergraduate teaching. We asked survey respondents to identify departmental colleagues who had facilitated change in their views about teaching and/or their teaching practice. Respondents reported experiencing change as a result of collegial interactions less often than they reported receiving resources from these interactions (Table 6). More than half (57%) of the faculty in the departments we studied were not reported to have caused any survey respondents to change their teaching views or practices.

After controlling for department and whether or not an individual was a survey respondent, several characteristics were significantly associated with how many times a faculty member was reported to have prompted a colleague to change his or her views or teaching practices: position type (i.e., DBER vs. non-DBER), participation in Regional Teaching PD, gender, academic rank, and how many resources the faculty member was reported to have provided to departmental colleagues (Table 7). DBER faculty were reported to have prompted almost three times as many occurrences of change in teaching views or practices than non-DBER faculty, and faculty who had participated in the Regional Teaching PD had prompted twice as many change occurrences as faculty who had not participated in the program (Table 7). Male faculty and full professors were reported to cause more change than their counterparts (Table 7). The number of resources for teaching that a faculty member provided to colleagues was significantly and positively associated with the number of times colleagues reported changing as a result of interacting with that faculty member. This result confirms that resources were a correlate of change.

### Table 7. Results of generalized linear regression models

<table>
<thead>
<tr>
<th>Response (dependent) variable</th>
<th>Number of resources provided</th>
<th>Number of changes caused</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept for count model</td>
<td>2.82** (1.31, 5.94)</td>
<td>0.14** (0.06, 0.32)</td>
</tr>
<tr>
<td>Gender (male vs. female)</td>
<td>0.72 (0.42, 1.19)</td>
<td>1.85* (1.16, 3.04)</td>
</tr>
<tr>
<td>Academic rank (full vs. assistant)</td>
<td>1.70* (1.08, 2.62)</td>
<td>1.76* (1.03, 3.06)</td>
</tr>
<tr>
<td>Academic rank (associate vs. assistant)</td>
<td>2.64*** (1.57, 4.98)</td>
<td>1.12 (0.68, 1.88)</td>
</tr>
<tr>
<td>Position type (DBER vs. non-DBER)</td>
<td>2.09* (1.06, 4.10)</td>
<td>2.90*** (1.62, 5.24)</td>
</tr>
<tr>
<td>Teaching PD (participant vs. nonparticipant)</td>
<td>0.98 (0.5, 1.74)</td>
<td>2.03* (1.14, 3.60)</td>
</tr>
<tr>
<td>Survey participant (Y vs. N)</td>
<td>1.90** (1.23, 3.08)</td>
<td>1.43 (0.98, 2.12)</td>
</tr>
<tr>
<td>Total resources provided</td>
<td>NA</td>
<td>1.15*** (1.13, 1.19)</td>
</tr>
<tr>
<td>Department Yellow vs. Blue</td>
<td>1.17 (0.72, 2.12)</td>
<td>0.58* (0.36, 0.92)</td>
</tr>
<tr>
<td>Department Red vs. Blue</td>
<td>1.01 (0.45, 2.40)</td>
<td>0.72 (0.39, 1.30)</td>
</tr>
<tr>
<td>Department Green vs. Blue</td>
<td>1.09 (0.68, 1.89)</td>
<td>0.37* (0.20, 0.68)</td>
</tr>
<tr>
<td>Intercept for binary model</td>
<td>2.48*** (0.81, 3.68)</td>
<td>NA</td>
</tr>
</tbody>
</table>

*Estimates and 95% confidence intervals of the association between explanatory variables and reports of teaching resources provided and change caused as a result of interactions about undergraduate teaching with departmental colleagues.

*In these models, assistant professor and the Blue department are set as the reference levels of the academic rank variable and department variable, respectively. One can fit another model with different reference levels to determine the relationship between these levels and the response variables. We examined all of these. The only significant results were associate vs. full professors for resources provided: 0.65*(0.35, 1.11) and Red vs. Green for change caused: 1.95*(1.10, 3.49). Setting a different reference level does not change other results, except the estimate of the intercept.

*p < 0.05; ** p < 0.01; *** p < 0.001 (two-tailed tests).
garnered from social interactions are important for effecting change. Survey respondents did not cause more change than nonrespondents (Table 7), indicating that interest in teaching (as inferred from status as a survey respondent) was not the same as driving change in teaching. Being influential, then, is more than being highly invested in undergraduate teaching. Faculty from some departments were more likely to be reported to cause change than others, including faculty from the Yellow department when compared with the Blue department, and faculty from the Red department when compared with the Green department. No other differences among departments were significant.

By examining the number of reported change occurrences caused by each faculty member, we identified nine opinion leaders in three departments: five DBER faculty and four non-DBER faculty. Across departments, opinion leaders caused 48, 67, and 69% of the change reported in the Green, Yellow, and Red departments, respectively. These values seem especially high when considering that opinion leaders are <15% of the faculty in each department. These results reveal who influences colleagues to change their teaching but provide little insight as to why these individuals are influential or how they are impacting their colleagues, which is addressed by our qualitative investigation.

**DBER versus Non-DBER Opinion Leaders**

Results of the qualitative analysis showed that DBER and non-DBER opinion leaders derived their credibility and reputation among colleagues differently. Non-DBER opinion leaders, without exception, had a reputation for being excellent teachers. These faculty were noted by their colleagues for the positive influence they have on undergraduates:

“I never saw anybody’s teaching evaluations so off scale, I mean, [Nathan] must be doing something wonderful in his classroom, but it wouldn’t surprise me because he is a very enthusiastic guy.”

They were also described as exceptionally dedicated instructors, as described below:

“I would say if there was one person that I think is an astonishing teacher and puts more effort into it than anyone else, I would say it would be Natalie. She just has a natural gift.”

Most of the non-DBER opinion leaders had many years to build their reputations within their departments. The median number of years they had been members of their departments was 21 (SD = 17). One of the four non-DBER opinion leaders had been active in the department for more than 40 yr and another for almost 30. Neither of these two faculty had had active research labs in many years. Despite long careers at the institution, some non-DBER opinion leaders had not been promoted beyond associate professors. This may be because their current positions emphasized teaching and service, which are not as highly valued in their departments as research, as illustrated by this quote:

“In [my department] it’s publish or perish. You either publish and get grants or you do not get tenure. And if you have tenure, and this is the case of one person in the department, if you have tenure but you’re not able to publish or get a significant number of grants, you’ll never become a full professor. The chair won’t even put you up for it. But this person does more teaching than anyone else in that department.”

The most junior non-DBER opinion leader had a typical teaching load for research-active faculty. He had recently been promoted, on schedule, to associate professor and was perceived by colleagues to be both an accomplished researcher and teacher, as described by a colleague:

“Nathan is a great example: his recognition, he has been recognized as a very, very good teacher … I think what he has done in his innovation of teaching … is exemplary, and he embraced it as a young research-active assistant professor at the highest level.”

Whereas all four of the non-DBER opinion leaders were described as excellent teachers, this was only true for two of the five DBER opinion leaders. That does not mean that DBER faculty were not perceived as great teachers but that their most salient role in their departments was not as a teacher. DBER faculty were perceived as credible because their area of professional expertise was education while also being experts in their biology disciplines, as illustrated by the following quote:

“[Diana] had a lot of credibility because of her [biology discipline] background. She knows [her biology discipline] inside and out and she knows instruction, she knows the techniques and she obviously enjoys assessing and understanding how well it’s working.”

DBER faculty were also seen as innovative educators with knowledge of teaching strategies that other faculty lacked. They were perceived to be the resident education experts who were willing and able to provide resources for teaching and access to other education experts. They were commonly described as the person faculty would seek out if they were interested in changing their teaching, as illustrated by the following quotes:

“Because of her willingness to discuss learning techniques or teaching techniques with us, providing us new materials, introducing sort of, for example, like the Top Hat system for classroom engagement and bringing back information from workshops. She has been really influential and it is my idea of what an educator can be.”

“Darcy has been influential just because everybody knows and respects her as being an outstanding teacher and one of the first to bring the active learning. So if there’s any question, she’s one of the first people anybody would ask.”

In comparison with non-DBER opinion leaders, DBER faculty tended to be newer to their positions. One DBER faculty member was a full professor, two were associate professors, and two were assistant professors. The median number of years they had held positions in their departments was 7 yr (SD = 6). All DBER faculty had positions with research expectations, but they also had higher teaching EFT (equivalent
full time) than other faculty in their departments, which means a greater percentage of their time was expected to be dedicated to teaching.

**Change in Undergraduate Teaching**

DBER and non-DBER opinion leaders were reported to contribute differently to changing teaching views and practices in the department. Although four non-DBER faculty were identified as opinion leaders using survey data, interview data did not confirm that these faculty influenced their colleagues to change their views and behaviors about teaching. Interview participants described only one of the four non-DBER faculty as prompting change. In contrast, interviewees described all five DBER opinion leaders as engendering change among departmental colleagues. Therefore, most of the following results pertain to DBER opinion leaders.

Our analysis revealed notable differences in the role and impact of opinion leaders across the Red, Yellow, and Green departments, so we present these results separately. Each section begins with one faculty member’s response to a question about what in his or her department helps him or her to be a good teacher. We selected these quotes to be representative of the most common ideas in each department and to introduce major themes observed for each department. At the end of this section, we summarize some potentially relevant differences in DBER positions among the three departments.

**Yellow Department**

*Interviewer:* What is going on in your department that might help you be a good teacher?

*Interviewee:* Well, I think the biggest thing that’s happened since I’ve gotten here is there has been an emphasis, I think university-wide, on bringing in science educators1 ... I have a natural draw toward the science educators that are here ... Since I came in as kind of an empty vessel as an instructor, the only baseline I had was the way I was instructed. Once I started interacting with Diana, and then went through the interview process and got to know Dakota ... I started thinking about other ways that you could instruct. These people that I mentioned ... were interested in teaching, instructing, using different methods other than lecturing. So I kind of got hooked into that way of instructing. I’m not there yet. I’m not as polished as I want to be, but especially considering where I am in my career, I’d really like to be able to say that by the time I retire, I could feel like I’m a really good instructor because I like doing it this way, I like interacting with the students, and I don’t particularly like standing up there and droning on and on and on, class after class.

The Yellow department had four opinion leaders according to survey data: two DBER and two non-DBER faculty. Both non-DBER opinion leaders were reputed to be dedicated educators who personally contributed substantially to undergraduate teaching in the department. However, neither non-DBER opinion leader was described by interviewees as prompting changes in colleagues’ views about teaching or teaching practices. The DBER opinion leaders were relatively new to the department, each having joined within the previous 3 yr, and each was perceived as having had a substantial impact on views about teaching in the department and the teaching practices of some of their colleagues.

The changes the DBER opinion leaders had engendered stemmed primarily from their roles as coordinators of a large course and through coteaching in that course with colleagues. The course they coordinated was the largest undergraduate course provided by the department, involving seven or more faculty per year, far more than any other single course. A small part of the DBER opinion leaders’ EFT was dedicated to the administration of this course. Before the arrival of the DBER opinion leaders, the course lacked explicit coordination, and faculty reported that the instructional quality had varied considerably. Course coordination completed by the DBER faculty included organizing meetings of course faculty to discuss common learning objectives and generally serving as a resource for instructors in the course. The DBER opinion leaders also designed and taught sections of the course that integrated evidence-based teaching strategies, freely shared all of their materials, and welcomed colleagues to observe their classes.

Faculty teaching in the coordinated course were encouraged, but not required, to be involved in identifying common goals and teaching strategies for the course. Colleagues appreciated that DBER faculty were welcoming, but not pushy. The general perception in the department was that no one should be forced to change his or her teaching but that support would be available to interested instructors. Some instructors of the coordinated course chose to collaborate with one another and others declined to do so, as described here:

“I think [Dakota’s] had to tiptoe around some people’s egos about how to instruct the course. I think there was a split in the instructors among just kind of ‘leave me alone’ and ‘I’m going to do my thing’ to the instructors that would attend meetings and get involved in unifying the material and some of the decisions about the way the course runs and kind of get on the same page. So I have attended those meetings, and Dakota has tried to develop kind of a unified curriculum.”

Collaborating with DBER opinion leaders prompted some faculty to critically examine their teaching practices. One interviewee reported that DBER faculty brought teaching to the spotlight in the department, making some faculty realize that “maybe we aren’t doing it as wonderfully well as we think we are.” Another interviewee described the impact of DBER faculty this way:

“I think now it’s different. I think there’s discussion about how you present the material ... Faculty members get all grumpy, because you’ve got to have more content and the [DBER faculty] are saying, ‘Well, yeah but you got to have—you’ve got to teach it right.’”

Faculty in the Yellow department reported that coteaching with DBER opinion leaders had led some faculty to change their teaching practices. Commonly, when two science faculty members teach in the same section of the same course, they teach mostly independently, one after the other. Coordination is often limited to dividing up topics to cover
and agreeing how grades will be calculated. Faculty who had cotaught with DBER opinion leaders reported having a much different experience. They described working closely with their DBER coteachers on both the content and teaching strategies used in the course. Coteachers regularly attended one another’s classes to observe and to help facilitate group work. Generally, this was the non-DBER faculty member’s first exposure to interactive strategies implemented in large courses, and the DBER opinion leader provided class structure and guidance to facilitate the success of these strategies. Interview respondents noted substantial changes in their colleagues’ teaching as a result of coteaching with DBER faculty, as described by this interviewee:

“I am really amazed that some of my colleagues whose teaching has been, like, radically transformed. I’m so surprised that they are, instead of just standing there lecturing, they are willing to break things up into small groups and wander around and kind of lose control of the classroom temporarily while the students do their own thing … So they’ll coteach something with Diana or with Dakota and suddenly they’ve been tossed out of this comfort zone they’ve been in for so many years and I have been really proud that now they are flexible enough to give it a shot … I know I have talked to a couple of them who said, ‘You know, I’m really enjoying this.’”

In addition to working with colleagues on teaching, the DBER opinion leaders in the Yellow department had collaborated on education components of grants for external funding. Collaborating with DBER faculty on proposals made some faculty more aware of education as an area of scholarship. The following quote describes how a faculty member more seriously considers the education components of his grants after his experience working with a DBER colleague:

“So, Diana recently has had a big impact in figuring out how to more effectively include educational opportunities in my research program. Seeing her commitment to evaluating how well an outreach activity—what impact it’s having on education—has kind of stepped up my commitment to taking that seriously.”

The addition of DBER faculty has prompted substantial instructional reform among some faculty in the Yellow department, but this was mostly limited to faculty with whom DBERs have cotaught. Some faculty in the coordinated course have opted not to collaborate with their peers and have likely continued to teach primarily using lectures. Additionally, an instructor who does not teach the coordinated course reported that support for teaching is still lacking in the department:

“I know [support in the coordinated course] really changed a lot when Diana and Dakota came in … But that hasn’t really filtered into our course at all. So I would just say the resources [for teaching] I mean it’s basically been a peer thing where if you have questions, you go talk to somebody who’s been teaching for a long time and you get sort of personal advice and maybe a syllabus or something … But as far as really formal sessions where people get together and compare notes and things like that, that’s just never been part of the department so far.”

Red Department

Interviewer: What might be going on in your department that helps you be a good teacher in general?
Interviewer: There is a reasonably strong emphasis on—they don’t ask us to do a whole lot of teaching, but I think the expectation then is for what little they ask us to do we ought to do it well. The general ethos of being aware that there is kind of a body of literature on how to teach well is useful. So partly, just because we are aware of it; partly because I feel like if I try and do something that might … not work the first time I try it, that I will be supported in that, because everyone else—enough of the department is aware of what I am trying to do and why I would try to do it that way. And then just having that network of people who are also trying to do this; some of whom who have been doing it for a long time; some of whom who have been teaching for a long time but are still relatively new to, are trying to change how they teach. So I have senior faculty who are trying to incorporate these active-learning strategies and they are as new in it as I am, but then I have got folks like Dawn and Darcy who are pretty [much] masters at it, so its useful to have that. It has been very helpful.

The Red department had three opinion leaders: two DBER and one non-DBER faculty, and all three were reported to have engendered change in the department. Interviewees most often described the role of the opinion leaders as contributing to the growth and maintenance of a culture that values undergraduate teaching, as illustrated by the following quote:

“I think the culture here is that we talk about [teaching], and we have people like Dawn and Darcy in our department, who actually, I think they educate the professors here in terms of how students learn. So in general, we just have this sort of atmosphere that people do talk about teaching. People do think it’s important. I think the value is sometimes it’s—that the community or the society sort of impose on people.”

The idea that particular viewpoints about teaching may be “imposed” on members of the Red department is consistent with this department’s high transitivity (Table 4). Ideas are reinforced by high transitivity, because actors receive information directly and indirectly. For example, in a two-legged triad (Figure 1), actor B can only receive information directly from actor A, whereas in a transitive triad, actor B is also connected to actor C and can therefore receive the same information indirectly through actor C. If actor A is an opinion leader promoting evidence-based teaching strategies, then actor B will be more likely to be exposed to these strategies through multiple channels in the Red department compared with other departments. Change may therefore spread more efficiently in the Red department. The perception that the Red department has a culture that is uniquely supportive of undergraduate teaching is corroborated by the fact that the Red department had the densest departmental network and survey respondents in this department reported interacting with a greater percentage of their colleagues about teaching (Table 4).

Opinion leaders in the Red department had invested considerable time in providing training and mentoring on
teaching for their colleagues, and these experiences had convinced interviewees that teaching was valued in the department. Many faculty in the Red department participated in Regional Teaching PD, in which the DBER opinion leaders had leadership roles. The department head strongly encouraged faculty to participate, which communicated that learning to teach well was an expectation in the department. Participation in a shared training experience was perceived to contribute to the culture of valuing teaching, as described here:

“I think probably half of our faculty has gone through the [Regional Teaching PD]. So there is a certain kind of critical mass effect that happens, I think, once the majority of people have been exposed to a certain way of thinking. It means you can talk to other people in that language.”

Faculty from other departments could also participate in Regional Teaching PD, but it was much less common. The Red department head advocated for faculty to participate, while the other department heads did not. It may be that the Red department head especially valued Regional Teaching PD because it was run by a DBER faculty member from the Red department.

The non-DBER opinion leader also provided mentoring that convinced colleagues teaching was valued. In fact, in more than one case, he had observed a colleagues’ class for most or all of a semester. The feedback he gave was formative and constructive in nature, rather than evaluative, shaping colleagues’ teaching, as one faculty member explained,

“So for example, I mean, you’ve probably heard many stories about Nicholas, but the first time I was teaching [introductory biology] was just a panic for me, because I’d never done that material and never done that size class, and Nicholas’ office is literally right next to mine. I didn’t ask him to do it and he didn’t tell me he was formally doing it, but for the weeks before hand and for the time I was teaching, I was here every day, every night, focusing on that, he would be here. And so I would constantly run over and then he’d get me three papers and I’d come back, and he’d give me this and I knew he was making himself available … I don’t know what we’ll do when he is not available to help with that.”

One interviewee had so valued the mentorship she received from this opinion leader that she had gone on to provide similar mentorship for another colleague, thus perpetuating actions that promoted quality teaching in the department.

Red department opinion leaders also contributed to a culture of valuing teaching by raising awareness about education research. They invited education experts to present at weekly departmental seminars, which were widely attended by departmental faculty. They had also presented their own education research to colleagues in regularly scheduled departmental gatherings, such as annual retreats. The current department head facilitated this by inviting the DBER faculty to present. Some interview participants indicated they would be unlikely to seek such experiences on their own, as illustrated by this quote:

“So, I think it was very interesting when we had, for example, a retreat. I recall maybe two or three years ago Dawn gave a very, very interesting talk about her research with the undergrads. We felt very, very intrigued to hear that. I mean typically, I will never go and hear those kinds of talks. But … I thought it was great.”

By providing opportunities to think and learn about teaching in standard departmental venues, the opinion leaders may have influenced colleagues who were unlikely to be influenced by efforts external to the department. The opinion leaders were also regularly vocal in department meetings about needing to provide quality undergraduate education, as illustrated in this quote:

“So, part of the role of people like Nicholas in the past has been to remind us of the importance of teaching and make sure it doesn’t get forgotten.”

Exposure to education research and to professional development that emphasized a scientific approach to teaching influenced how faculty in the Red department thought about teaching. Specifically, many faculty shared the perception that teaching should be informed by education research. Some interviewees even reported that there is an understood expectation in the department that faculty will approach their teaching scientifically, as described in this quote:

“And I also feel like, partly because of Darcy and Dawn’s presence that has helped, and still does, if you are going to teach, you should do it well, and you should think scientifically about how to do it well.”

Only a few interviewees in the Red department reported that they had changed their teaching practice as a direct result of interacting with opinion leaders. However, the culture of valuing both teaching and education research had encouraged several faculty to incorporate recommended strategies in their teaching. The following quote emphasizes that changes in teaching practice may only occur after multiple encounters with a new strategy:

“I didn’t adopt using student response systems in the large class until I had actually seen Darcy do a presentation at one of our faculty retreats and then I had seen another presentation and then I had gone to [Regional Teaching PD]. So I actually was exposed to it multiple times through the department before I finally made the leap. So having those resources where they can kind of keep you up-to-date, knowing there are any number of people here that I could have a conversation with about it, as I said we are bringing in speakers, so sort of having that support network.”

Green Department

Interviewer: What might be going on your department that helps you be a good teacher?
Interviewer: I think having Danielle in the department and other people in the department who are interested in teaching—there have been conversations about flipping classes. And so I think it’s made the rest of
The Green department was the smallest department we investigated, and members of the department described it as very “collegial.” Two interviewees referred to the department as a “family.” Network density in this department was high compared with the Blue and Yellow departments, meaning that more possible ties existed (Table 4). However, transitivity was lower in this department than in the Red or Yellow department (Quardonokus and Henderson, 2015). Low transitivity slows the spread of ideas, allowing differing views and behaviors to persist and potentially hindering reform.

The Green department had one DBER and one non-DBER opinion leader. The non-DBER faculty member was widely respected as a superb teacher but was not perceived by interviewees as prompting colleagues to change their views about teaching or their teaching practices. The DBER opinion leader was credited with teaching a highly innovative course in the department and raising awareness about evidence-based teaching strategies among colleagues, including some strategies that other faculty tried in their own classes. The DBER opinion leader raised awareness by presenting about her teaching practices at a monthly faculty gathering that was normally dedicated to discussion about research. This presentation was well received by colleagues, and as a result, two other faculty members had presented or planned to present and get feedback about their teaching. This represented a shift in the department, because teaching had become an acceptable discussion topic in a venue normally reserved for science research discussions. The DBER opinion leader had also invited education speakers to the weekly departmental seminar:

“[Danielle] brought in seminar speakers who were in the area of education and they were part of the departmental seminar series that we all went to. So that broadened, I think, everybody’s horizons, including mine a little bit, to hear those seminars that were more from field of education or science education than from the normal [biology] research talk.”

Moreover, the DBER opinion leader had a reputation as always being available to faculty for any education question or challenge and always willing to help:

“[Danielle] has been really influential because of her willingness to discuss learning techniques or teaching techniques with us, providing us new course room materials, introducing … the [personal response system] for classroom engagement and bringing back information from workshops. She has been really influential and it is my idea of what an educator can be.”

Few interviewees in the Green department talked about changing their teaching views and practices as a result of interacting with colleagues. One person saw academics as people who “do what they want to do … and run our courses the way we want to run them,” indicating a limited influence of any single faculty member:

“Danielle has introduced people to some new concepts but I don’t know how much people are really influenced to change by a particular person.”

Another faculty member felt that she valued teaching more than many people in her department and that only a few colleagues cared about teaching:

“So I think that, for example, Natalie really respects what I do as far as teaching and I really respect what she does as far as teaching, but I think some of the acknowledgement or understanding is kind of limited to a small number of people.”

Notably, interviewees from the Green department discussed how DBER faculty from other departments had provided them with resources and even prompted them to change their teaching. DBER opinion leaders from other departments had spoken at a weekly seminar in the Green department, provided teaching observations for multiple members of the department, and cotaught with faculty from the Green department, potentially making other DBER opinion leaders as or more influential than the opinion leaders within the department.

**Differences among Departments in DBER Positions**

The roles, responsibilities, and histories of DBER positions within the three departments could be influencing the degree to which DBER faculty engender change in their departments. DBER faculty in the Yellow department had formal leadership roles for teaching in the department, because they coordinated a course that included many faculty. DBER faculty in the other two departments did not hold this role. Yellow department DBER faculty regularly cotaught and taught a course their colleagues considered important, while Red and Green department DBER faculty taught classes less valued by their colleagues. In fact, DBER faculty in the Red department rarely taught in their own department, because they were responsible for introductory biology courses that serve many departments.

The history of DBER appointments was also different in the three departments. A point of pride for interviewees in the Red department was that they had a long history of employing faculty who focus on teaching. Indeed, they were the first life sciences department at this institution to have a DBER faculty member. However, for many years, the teaching-focused faculty in the department did not engage in research and were considered by many to be less impressive than research faculty. The current DBERs are somewhat saddled with the legacy of these undervalued teachers, because faculty who teach but do not conduct research and DBERs are grouped together in the minds of many faculty in the department. In contrast, the Yellow department had only recently hired DBER faculty, but the DBER positions in this department began as positions with research expectations, and the department was interested in building a national reputation as a home for influential biology education research. In the
Green department, the DBER faculty member was the first in this type of position and had different experiences than DBERs in other departments. The responsibilities of the DBER position had changed considerably in a short time frame and included extensive service to the department. The DBER faculty member in this department left the position shortly after these data were collected.

DISCUSSION

The idea that talking about teaching with colleagues can influence change is not new (e.g., Palmer, 1993), but little empirical evidence specifically supports the idea that colleague–colleague interactions can support reform in college science teaching. Our results demonstrate that, at least within one research institution, life sciences faculty receive social resources (e.g., support, information, feedback) from their colleagues and that this exchange of resources is associated with changes in teaching views and practices. Overall, interactions about undergraduate teaching were not common (e.g., more than half of respondents reported interacting about teaching with no more than one colleague per month; Table 4), leaving ample room to increase social interactions as a strategy for promoting education reform.

Who Promotes Teaching Change?

In this study, DBER faculty were perceived as particularly influential to undergraduate teaching, and faculty who engaged in the Regional Teaching PD had more impact on colleagues’ teaching than faculty without this training. DBER faculty and faculty who engage in teaching professional development are likely to have specialized knowledge of teaching and learning that other faculty lack, and this knowledge may enable them to engender change. These observations provide preliminary support for the first hypothesis generated by this exploratory study:

Hypothesis 1: Faculty who are perceived by departmental colleagues as having unique expertise in education and being accessible will foster more adoption of evidence-based teaching strategies than faculty without this reputation.

Borgatti and Cross (2003) propose that individuals seek information in their networks (i.e., departments) depending on their knowledge of what another individual knows, their value of that individual’s expertise, and their ability to access that person’s expertise in a timely manner (Borgatti and Cross, 2003; Cross and Borgatti, 2004). Results from this study provide empirical support that the characteristics of knowledge, value, and access are associated not only with the behavior of seeking information related to undergraduate teaching but also with taking action on that information.

DBER faculty are well situated to be hubs of teaching knowledge for departmental colleagues. Because DBER faculty hold positions in science departments and are hired specifically to study undergraduate science education, colleagues are likely to know that DBERs have expertise in undergraduate teaching. Faculty in this study valued the education expertise of DBER opinion leaders, and many saw that expertise as distinct from their own knowledge of teaching and learning. Furthermore, DBER faculty were perceived as accessible sources of information and resources about teaching. Similarly, faculty who had participated in Regional Teaching PD may have been perceived as having gained valuable, specialized knowledge and were embedded in the department and therefore easy to access. Participants in Regional Teaching PD may be more impactful to colleagues, because they are more likely to have ties with colleagues than do nonparticipants in this program. In contrast, DBER faculty were no more likely to have ties about undergraduate teaching than were non-DBER faculty, yet the change they were perceived to cause was much more extensive. This suggests that the influence of DBER faculty is not due to the degree to which they interact with colleagues about teaching, but how they interact.

Future research is necessary to determine how faculty perceive their colleagues’ knowledge of teaching, what teaching knowledge they particularly value, and what maximizes perceptions of accessibility. As this study indicates that DBER faculty are perceived as particularly impactful, this population may be a promising group in which to start asking these questions.

What Opportunities Promote Teaching Change?

In this study, three types of opportunities to interact with colleagues about teaching were perceived by faculty as influential to views about teaching and teaching practices, including coteaching, teacher training, and formal presentations of education scholarship. Commonly, faculty had these experiences as a result of department-level decisions rather than as a result of seeking these experiences on their own, providing preliminary support for the second hypothesis generated by this exploratory study:

Hypothesis 2: Department-level change will be necessary to promote colleague–colleague interactions about undergraduate teaching.

Coteaching, teacher training, and formal presentation of education scholarship changed thoughts about teaching and teaching practices, and each could be influenced by action at the department level. First, faculty who cotaught with DBER opinion leaders reported the most change in their attitudes and behaviors, corroborating other studies demonstrating the value of coteaching (e.g., Orlander et al., 2000; Wright, 2005; Henderson et al., 2009). The department influences coteaching because department leadership assigns teaching responsibilities. Coteaching may be a powerful strategy for engendering change, because it makes opinion leaders especially accessible to colleagues (access), while also providing the opportunity to learn about what opinion leaders know (knowledge) and to form a positive evaluation of their knowledge (value). DBER faculty were accessible to their coinstructors operationally, because they planned their courses together and attended each other’s class sessions, and intellectually, because they had to agree on how to solve a common problem (i.e., how to teach the course). Pairing DBER faculty or other reform-minded instructors to teach with new faculty could be a long-term plan for transforming teaching within departments, since new faculty have not yet invested time in developing teaching materials or established their own
identities as college instructors. Ultimately, protégés could become mentors of other faculty through coteaching. Like most interventions to promote reform, a coteaching model requires resources, primarily in the form of time, as the time and effort required for coteaching may be greater than for teaching half a course (e.g., Beach et al., 2008).

Second, teaching training was impactful for some faculty, and interactions with department heads and other departmental colleagues were key in promoting such experiences. For example, the head of the Red department encouraged faculty, sometimes forcefully, to attend Regional Teaching PD, and more than 10 people from the Red department attended, compared with less than half the faculty from any other department. A few teams from the other departments also participated in Regional Teaching PD, and each of these teams included one opinion leader (as designated by our survey results). In at least one case, an interviewee reported that he would not have participated if it were not for an opinion leader in his department making him aware of the program and inviting him to join a departmental team. Participation in Regional Teaching PD was especially notable, because it required a substantial time commitment—five full days—and because faculty in these departments rarely participate in other teaching professional development, even though the Center for Teaching and Learning and other groups on campus provide many different opportunities for teacher training.

Department heads can insist that new faculty or faculty who are struggling as teachers engage in teaching professional development or work with a mentor. This leverage is particularly powerful for pretenure faculty. Eventually, the department head may be able to play a smaller role in mandating participation as it becomes commonplace to seek these types of opportunities. This progression seemed to be underway within the Red department, because it was widely agreed by interviewees that all faculty teaching undergraduates should seriously consider participating in Regional Teaching PD.

The third type of experience that appeared to be an effective strategy for promoting teaching change was introducing education research and evidence-based teaching strategies at standard department events. Department heads or appointed committees within a department generally make arrangements for seminar speakers, faculty retreats, and other faculty gatherings and could therefore prioritize providing regular opportunities for faculty to be exposed to education scholarship. Although the DBER faculty in this study were able to present their own research, inviting outside speakers to departmental seminars also appeared fruitful. Faculty who have engaged in intensive teaching professional development could also be encouraged to formally share what they have learned. These events may be the only exposure many life sciences faculty get to the scholarship of education.

Limitations

This work is an in-depth case study of collegial interactions in life sciences departments at one research university, and our findings should not be inferred to other departments until additional studies have determined whether these patterns are observable in other departments and institutions. This is especially important when considering the findings about the influence of opinion leaders, especially DBER faculty. Colleague–colleague interactions take place within complex social structures with varied histories, people, and contingencies. For example, the personality and interests of a DBER faculty member could greatly impact the receptiveness of colleagues to his or her influence, as might the attitudes that non-DBER faculty have about the value and role of DBER faculty in a department. Indeed, in a large national survey, DBER faculty had diverse ideas about the most valuable contributions they could make in their departments (Bush et al., 2015). Large-scale studies of multiple departments across institutions are necessary before more generalizable inferences can be made.

ACKNOWLEDGMENTS

We thank the participating instructors for their time and willingness to share their experiences. Thank you also to the Biology Education Research Group at the University of Georgia, who improved the quality of this work with critical feedback, high standards, and endless support. A special thanks to Julie Stanton and Jennifer Thompson for reviewing our qualitative analysis and engaging in multiple fruitful discussions as this article evolved. Thank you to three anonymous reviewers and the editors for providing valuable feedback. Support for this work was provided by the University of Georgia.

REFERENCES

Am. Acad. for Advancement of Science (2011). Vision and Change in Undergraduate Biology Education: A Call to Action, Washington, DC.
Austin AE (2002). Preparing the next generation of faculty: graduate school as socialization to the academic career. J High Educ 73, 94–122.


A Campus-Wide Investigation of Clicker Implementation: The Status of Peer Discussion in STEM Classes

Justin D. Lewin,*† Erin L. Vinson,*† MacKenzie R. Stetzer,†‡ and Michelle K. Smith*†

*School of Biology and Ecology, University of Maine, Orono, ME 04469-5751; † Maine Center for Research in STEM Education, University of Maine, Orono, ME 04469-5727; ‡ Department of Physics and Astronomy, University of Maine, Orono, ME 04469-5709

Submitted October 26, 2015; Revised December 19, 2015; Accepted January 5, 2016
Monitoring Editor: Peggy Brickman

At the University of Maine, middle and high school teachers observed more than 250 university science, technology, engineering, and mathematics classes and collected information on the nature of instruction, including how clickers were being used. Comparisons of classes taught with \((n = 80)\) and without \((n = 184)\) clickers show that, while instructional behaviors differ, the use of clickers alone does not significantly impact the time instructors spend lecturing. One possible explanation stems from the observation of three distinct modes of clicker use: peer discussion, in which students had the opportunity to talk with one another during clicker questions; individual thinking, in which no peer discussion was observed; and alternative collaboration, in which students had time for discussion, but it was not paired with clicker questions. Investigation of these modes revealed differences in the range of behaviors, the amount of time instructors lecture, and how challenging the clicker questions were to answer. Because instructors can vary their instructional style from one clicker question to the next, we also explored differences in how individual instructors incorporated peer discussion during clicker questions. These findings provide new insights into the range of clicker implementation at a campus-wide level and how such findings can be used to inform targeted professional development for faculty.

INTRODUCTION

A number of national reports informed by emerging education research have advocated for active-engagement instruction in postsecondary science, technology, engineering, and mathematics (STEM) courses (American Association for the Advancement of Science, 2010; President’s Council of Advisors on Science and Technology [PCAST], 2012; Singer et al., 2012). Moreover, a recent comprehensive meta-analysis of 225 science education research articles indicates that students learn more in and are less likely to drop out of STEM courses that use these active-engagement instructional approaches (Freeman et al., 2014). One such instructional approach involves instructors posing multiple-choice conceptual questions, fostering peer discussion about these questions among the students, and asking students to indicate their answers via personal response systems or clickers.

Clickers are electronic voting devices that allow instructors to obtain real-time student responses to multiple-choice questions in order to assess student thinking and to inform instruction (e.g., Mazur, 1997; Caldwell, 2007; Smith et al., 2011). A recent nationwide survey found that 86% of U.S. college faculty members are familiar with clickers and 12% of faculty members have adopted clickers in their own classrooms (FTI Consulting, 2015). In addition, a study of student engagement in a large-enrollment undergraduate science class found the use of clicker questions and the follow-up to clicker questions to be the most engaging of all in-class activities observed, as measured by the Behavioral Engagement Related to Instruction protocol (Lane and Harris, 2015).

© 2016 J. D. Lewin et al. CBE—Life Sciences Education © 2016 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.
question; 2) students are given time to think; 3) students determine their individual answers and vote; 4) if there is variation in the student answers, neighboring students discuss their answers with one another; 5) students vote again after peer discussion; and 6) the instructor explains the correct answer to the whole class, often displaying a histogram of all student responses and soliciting explanations from students for incorrect and correct answers. Peer instruction encourages student interactions during lecture and breaks the monotony of passive listening while also offering an opportunity for the instructor to walk around the room and interact with students. These interactions allow the instructor to gauge the level of student understanding and thus gain insight into incorrect lines of reasoning.

Multiple studies have shown that the peer discussion portion of peer instruction increases student performance on clicker questions (Smith et al., 2009; Porter et al., 2011; Knight et al., 2015; Barth-Cohen et al., 2016) and that peer discussion produces higher performance outcomes when compared with other tasks such as quiet reflection (Lasry et al., 2009). Furthermore, the largest gains in student performance occur when peer discussion is immediately followed by an instructor explanation (Smith et al., 2011). Notably, students have more positive attitudes about the utility of clickers when faculty encourage peer discussion and are successfully able to create opportunities for students to discuss the multiple-choice questions (Keller et al., 2007).

Although the learning benefits of using clicker questions with peer discussion have been documented through the use of carefully designed protocols, it has also been noted that faculty often change and modify research-based pedagogies and tools, such as clickers, in their classrooms (Henderson and Dancy, 2007; National Research Council [NRC], 2013). For example, survey results from faculty in multiple disciplines showed that 15% of faculty who use clickers reported that they did not allow or did not encourage peer discussion during clicker questions (Keller et al., 2007). In another observation-based study of undergraduate physics classrooms, researchers found that none of the faculty had students record their individual answers before talking with peers (Turpen and Finkelstein, 2009), a component of peer instruction that is advocated by researchers (Mazur, 1997; Smith et al., 2011).

In this study, we used classroom observation data from 21 different STEM departments to explore the spectrum of instructional practices associated with clickers. Specifically, we asked: 1) Are there differences in instructional behaviors in classes that are taught with and without clicker questions? 2) In classes that use clickers, is there variation in how clickers are implemented? 3) How do individual instructors vary their implementation of clicker questions? The answers to all three questions are critical for identifying common-use cases on which to focus future research and for optimizing faculty professional development so that it may better support the effective implementation of clickers.

**METHODS**

For this investigation, University of Maine STEM instructors were sent emails asking them if they would allow middle and high school teachers to visit their classrooms and collect observation data. The instructors were receptive, with 74% agreeing to allow the teachers to observe their courses. Faculty who declined typically cited reasons such as giving an exam, canceling class, or having a guest lecturer present on the day of the proposed observation.

Observations were conducted in both February and April during the Spring 2014 and Spring 2015 semesters and in November during the Fall 2014 semester. Altogether, 270 class sessions were observed. These observations represented 119 instructors who taught 138 courses in 21 different departments (biology and ecology; chemical and biological engineering; chemistry; civil and environmental engineering; computer sciences; earth sciences; ecology and environmental sciences; economics; electrical and computer engineering; food and agriculture; forest resources; marine science; mathematics and statistics; mechanical engineering; molecular and biomedical science; new media; nursing; physics and astronomy; plant, soil, and environmental science; psychology; and wildlife, fisheries, and conservation biology). On average there were 12.6 (SE ± 2.5) class sessions observed per department. Demographic information about the types of courses and instructors observed is included in Figure 1. Data from 97 observations from Spring 2014 were reported in an earlier study (Smith et al., 2014).

All faculty members who agreed to be observed were given a human subjects consent form. Approval to evaluate teacher observation data of classrooms (exempt status, protocol no. 2013-02-06) was granted by the institutional review board at the University of Maine. Because of the delicate nature of sharing observation data with other faculty members and administrators, the consent form explained that the data would only be presented in aggregate and would not be subdivided according to variables such as department. Faculty members were given access to observation data from their own course(s) upon request after we collected observation...
and survey data for this study. In total, 68% of the observed faculty members requested their data and met with a professional development coordinator to discuss the results.

Selection and Training of Middle and High School Teachers

Thirty-eight teachers from the state of Maine conducted the classroom observations. To record instructional behaviors in the classroom, the middle and high school teachers were trained to use the Classroom Observation Protocol for Undergraduate STEM (COPUS) according to the training procedure outlined by Smith et al. (2013). COPUS was adapted from the Teaching Dimensions Observation Protocol (Hora et al., 2013; Hora and Ferrare, 2014). Briefly, at the beginning of the training, the 25 COPUS codes and code descriptions (Table 1) were discussed with the teachers (sample COPUS protocol sheets can be found in Smith et al., 2013, and at www.cwsei.ubc.ca/resources/COPUS.htm). The teachers then practiced coding videos of classrooms and discussed codes that were not unanimously selected. In total, the training took approximately 2 h. Teachers conducted observations in pairs, and we calculated Cohen’s kappa scores to measure interrater reliability (Landis and Koch, 1977) for every in-class observation; more details are included in the next section.

Analyzing COPUS Data

Pairs of teachers observed classes and were instructed to record their COPUS results independently. Cohen’s kappa interrater scores were calculated for each observation pair to establish coder reliability. The mean Cohen’s kappa was 0.89 (SE ± 0.006) for all observations. Because some of the paired observations had low interrater reliability scores, we removed the six lowest paired observations, which all had Cohen’s kappa interrater scores below 0.650. The average Cohen’s kappa interrater score for the remaining 264 observations was 0.91 (SE ± 0.005), indicating strong agreement among paired observers (Landis and Koch, 1977). With strong reliability between coders, only codes that both observers marked during each 2-min interval were included in the data set analyzed for this study.

To determine the relative abundance of each COPUS code, we added the total number of times each code was marked and divided by the total number of codes, resulting in a percent of code. For example, if both observers marked instructor lecturer (Lec) during the same 24 time intervals in a 50-min class period and marked 29 total instructor codes for the duration of the class, then 24/29 or 82.8% of the instructor codes correspond to lecture. Because it was difficult to visually represent and compare 25 COPUS codes in 264 different class sessions, we also collapsed the codes into four categories that describe what the instructor is doing and four categories that describe what the students are doing, as reported in Smith et al. (2014) and shown in Table 1. The percentage of each collapsed code was determined by adding the percent code of each individual code within the collapsed category. This collapsed-code representation is advantageous because it allows for a holistic view of multiple COPUS codes at the same time and facilitates comparisons across broad instructional approaches.

However, when trying to determine and compare the frequency of a single code, such as instructor lecturer (Lec) or student listening (L), percent-code calculations can be difficult to interpret, because multiple COPUS codes can be marked at the same time, which in turn can impact the denominator of the calculation (Lund et al., 2015). In particular, some codes are often marked together, such as instructor real-time writing (RTW) and instructor lecturing (Lec). Therefore, we also compared class sessions by calculating the percentage of 2-min time intervals in which specific codes, such as instructor lecturing (Lec) or student listening (L), were observed. The percentage of 2-min time intervals was determined by counting the number of 2-min time intervals in which each code was marked and then dividing that by the total number of time intervals that were coded. For example, if instructor lecturer (Lec) was marked during 24 two-minute time intervals out of the possible 25 two-minute time intervals then 24/25 or 96.0% of the possible 2-min time intervals contained lecture.

Table 1. COPUS instrument codes used to describe instructor and student behaviors in class and a description of the collapsed COPUS codes used to compare class sessions

<table>
<thead>
<tr>
<th>COPUS codes</th>
<th>Collapsed codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lec—lecturing or presenting information</td>
<td>Presenting</td>
</tr>
<tr>
<td>RTW—real-time writing</td>
<td></td>
</tr>
<tr>
<td>D/V—showing or conducting a demo, experiment, simulation, etc.</td>
<td></td>
</tr>
<tr>
<td>F/U—follow-up/feedback on clicker question or activity to class</td>
<td></td>
</tr>
<tr>
<td>PQ—posing nonclicker question to students (nonrhetorical)</td>
<td>Guiding</td>
</tr>
<tr>
<td>CQ—asking clicker question (entire time, not just when first asked)</td>
<td></td>
</tr>
<tr>
<td>AnQ—listening to and answering student questions to entire class</td>
<td></td>
</tr>
<tr>
<td>MG—moving through class guiding ongoing student work</td>
<td></td>
</tr>
<tr>
<td>1o1—one-on-one extended discussion with individual students</td>
<td></td>
</tr>
<tr>
<td>Adm—administration (assign homework, return tests, etc.)</td>
<td>Administration</td>
</tr>
<tr>
<td>W—waiting (instructor late, working on fixing AV problems, etc.)</td>
<td>Other</td>
</tr>
<tr>
<td>O—other</td>
<td></td>
</tr>
<tr>
<td>L—listening to instructor</td>
<td>Receiving</td>
</tr>
<tr>
<td>Ind—individual thinking/problem solving</td>
<td></td>
</tr>
<tr>
<td>CG—discussing clicker question in groups of two or more students</td>
<td></td>
</tr>
<tr>
<td>WG—working in groups on worksheet activity</td>
<td></td>
</tr>
<tr>
<td>OG—other assigned group activity</td>
<td>Students working</td>
</tr>
<tr>
<td>Prd—making a prediction about a demo or experiment</td>
<td></td>
</tr>
<tr>
<td>TQ—test or quiz</td>
<td></td>
</tr>
<tr>
<td>AnQ—student answering question posed by instructor</td>
<td>Students talking to class</td>
</tr>
<tr>
<td>SQ—student asks question</td>
<td></td>
</tr>
<tr>
<td>WC—students engaged in whole-class discussion</td>
<td></td>
</tr>
<tr>
<td>SP—students presenting to entire class</td>
<td></td>
</tr>
<tr>
<td>W—waiting (instructor late, working on fixing AV problems, etc.)</td>
<td>Other</td>
</tr>
<tr>
<td>O—other</td>
<td></td>
</tr>
</tbody>
</table>
Figure 2. Example excerpt of COPUS codes from a clicker class session with three clicker episodes. Abbreviated COPUS codes, described in Table 1, are along the top; the 2-min time intervals are along the left side. Student and instructor codes of interest are shaded red, with blue boxes surrounding each clicker episode.

Analyzing Instances of Clicker Use
For this study, we were particularly interested in class sessions that used clicker questions. At the University of Maine, instructors largely started implementing clickers in their classrooms in 2005 (Strukov, 2008), and the Faculty Development Center estimates that currently more than 5000 students are enrolled in courses that use clickers each academic year. To determine which class sessions used clickers, we looked for instances in which the instructor clicker question (CQ) code was marked by both observers during a single 2-min interval. The CQ code was identified in 80 of the 264 class sessions observed.

To find out more about how the clicker questions were used in each of the 80 class sessions, we looked for blocks of 2-min time intervals with individual or consecutive instructor clicker question (CQ) codes, and we called these “clicker episodes” (Figure 2). In total, 181 clicker episodes were observed, and the duration of each episode was determined by counting the number of consecutive 2-min time intervals marked with clicker question (CQ) codes. Overall, clicker episodes had a mean duration of 2.4 (SE ± 0.11) 2-min time intervals; therefore, the average clicker episode was less than 5 min (2 × 2.4 = 4.8) in duration. To determine how clicker questions were used during these episodes, we examined student behaviors during the same time intervals. Only two student behaviors were selected along with instructor CQ: individual thinking (Ind) and clicker group discussion (CG).

Thus, there were three possible combinations of student behaviors during clicker episodes: Individual Thinking Only, Peer Discussion Only, and Individual Thinking and Peer Discussion Combined. For example, Figure 2 shows a class session with three clicker episodes, indicated by three segments of time with instructor clicker question (CQ) codes that are separated by one or more 2-min time intervals. The first clicker episode is solely characterized by individual thinking (Ind) student codes and is classified as Individual Thinking Only. The second clicker episode has clicker group work (CG) only and is classified as Peer Discussion Only. The third clicker episode has individual thinking (Ind) student codes followed by clicker group discussion (CG) codes, and is classified as Individual Thinking and Peer Discussion Combined.

Clicker Use in Class Sessions
Upon analysis of all clicker episodes, two broad class session modes were identified, those with peer discussion during clicker questions and those without peer discussion during clicker questions (Figure 3). Peer Discussion class sessions had at least one clicker episode with peer discussion, indicated by a clicker group discussion (CG) code with a corresponding instructor clicker question (CQ) code (Figure 4A). These class sessions may have also included student individual (Ind) or group work (WG, OG) codes, but the presence of at least one student CG code defined this mode.

The class sessions without peer discussion during clicker questions were further classified into two modes, Individual Thinking and Alternative Collaboration (Figure 3). Individual Thinking class sessions had no clicker episodes with peer discussion, and were thus characterized by the presence of instructor clicker question (CQ) codes paired with student individual thinking (Ind) codes (Figure 4B). In the Individual Thinking class sessions, students never discussed class material in groups. Alternative Collaboration class sessions had no clicker episodes with peer discussion, and therefore individual thinking (Ind) was the only student code that coincided with instructor clicker question (CQ) codes. However, Alternative Collaboration class sessions included worksheet-based group work (WG) or other group work (OG) at another point in the class period (Figure 4C). While students in Alternative Collaboration class sessions voted on clicker questions as individuals, just like students in Individual Thinking class sessions, observer notes suggested that the clicker questions were often tied to the group activities. Thus, students in some Alternative Collaboration class sessions had the opportunity to discuss relevant question material with peers, just not in the context of the clicker questions themselves.

To determine whether there is variation in how individual instructors use clickers at different times, we looked at clicker data for instructors who were observed teaching with at least two clicker episodes (n = 25 instructors). Specifically, we examined the percentage of clicker questions that had Individual Thinking Only, Peer Discussion Only, and Individual Thinking and Peer Discussion Combined. The percent was calculated by dividing the number of episodes in each category by the sum of all episodes for that instructor.

Observer Feedback
To collect additional information about class sessions and to give the middle and high school teachers opportunities to reflect on the instruction, we developed a feedback survey for
the teachers to complete after each observation. The survey was developed during Summer 2014, piloted in Fall 2014, revised based on teacher interviews and written feedback, and implemented in Spring 2015. Observers completed the survey for each observation in pairs, discussing their reasoning for each answer. Discussions were audio recorded to monitor the usefulness of the survey and to ascertain whether or not additional clarification was needed on any items.

### A. Peer Discussion Class Session

<table>
<thead>
<tr>
<th>Time</th>
<th>L</th>
<th>Ind</th>
<th>CG</th>
<th>WG</th>
<th>OG</th>
<th>AnQ</th>
<th>SQ</th>
<th>WC</th>
<th>Prd</th>
<th>SP</th>
<th>T/Q</th>
<th>W</th>
<th>O</th>
<th>Lec</th>
<th>RW</th>
<th>FlI/p</th>
<th>PQ</th>
<th>CQ</th>
<th>AnQ</th>
<th>MG</th>
<th>1o1</th>
<th>D/V</th>
<th>Adm</th>
<th>W</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 - 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 - 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 - 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 - 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### B. Individual Thinking Class Session

<table>
<thead>
<tr>
<th>Time</th>
<th>L</th>
<th>Ind</th>
<th>CG</th>
<th>WG</th>
<th>OG</th>
<th>AnQ</th>
<th>SQ</th>
<th>WC</th>
<th>Prd</th>
<th>SP</th>
<th>T/Q</th>
<th>W</th>
<th>O</th>
<th>Lec</th>
<th>RW</th>
<th>FlI/p</th>
<th>PQ</th>
<th>CQ</th>
<th>AnQ</th>
<th>MG</th>
<th>1o1</th>
<th>D/V</th>
<th>Adm</th>
<th>W</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 - 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 - 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 - 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 - 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### C. Alternative Collaboration Class Session

<table>
<thead>
<tr>
<th>Time</th>
<th>L</th>
<th>Ind</th>
<th>CG</th>
<th>WG</th>
<th>OG</th>
<th>AnQ</th>
<th>SQ</th>
<th>WC</th>
<th>Prd</th>
<th>SP</th>
<th>T/Q</th>
<th>W</th>
<th>O</th>
<th>Lec</th>
<th>RW</th>
<th>FlI/p</th>
<th>PQ</th>
<th>CQ</th>
<th>AnQ</th>
<th>MG</th>
<th>1o1</th>
<th>D/V</th>
<th>Adm</th>
<th>W</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 - 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 - 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 - 8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 - 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3. Description of the three distinct modes of clicker class sessions: Peer Discussion, Individual Thinking, and Alternative Collaboration. These modes were identified based on the presence of an instructor clicker question (CQ) code and four student codes: individual thinking (Ind), clicker group discussion (CG), worksheet group work (WG), and other group work (OG).

Figure 4. Example excerpt of COPUS codes for the three class modes: (A) Peer Discussion class sessions, which include the presence of at least one student clicker group discussion (CG) code. (B) Individual Thinking class sessions, in which students did not have the opportunity to talk in groups, so only the student individual thinking (Ind) code was selected. (C) Alternative Collaboration class sessions, in which students discussed material in groups (student codes worksheet group work [WG] and/or other group work [OG]) but did not have peer discussion during the clicker question. Abbreviated COPUS codes, described in Table 1, are listed along the top; the 2-min time intervals are indicated along the left side. Student and instructor codes of interest are shaded red.
RESULTS

Characterizing Instructional Behaviors in Class Sessions Taught with and without Clicker Questions

To determine whether there are differences in instructional behaviors between classes taught with and without clickers, we first separated out class sessions that included clicker questions. Eighty of the 264 observed class sessions featured at least one clicker question (CQ) instructor code and were thus classified as clicker class sessions (Figure 3). We examined differences in class size and found that the average student enrollment in classes that use clickers (111 students) was significantly higher than that in classes that did not use clickers (72 students) (independent-samples t-test, \( p < 0.05 \)).

To have a holistic view of the COPUS codes at the same time in both clicker and nonclicker classes, we examined the distribution of collapsed-code percentages and found a range of classroom behaviors (Figure 5). Notably, the non-clicker class sessions range from 0 to 100% Instructor Presenting and Student Receiving, whereas the clicker class sessions have a narrower range from 13 to 87% Instructor Presenting and 19 to 89% Student Receiving.

A section of this survey included a question about how challenging the clicker questions were for students. In general, the clicker questions were:

1. Challenging for students—the class vote was often split.
2. Easy for students—the majority of the students answered correctly.
3. Cannot determine—the instructor did not talk about the class voting results.

To help teachers answer this question, during training sessions, we watched training videos of instructors using clicker questions and showed teachers how the clicker system worked in detail by asking them a few clicker questions. We also demonstrated how the clicker results could be displayed to students and how the teachers could learn about how challenging the questions are for students based on the voting results. Because this survey was not fully implemented until Spring 2015, feedback is provided for 41 of the 80 clicker class sessions. Seventeen middle and high school teachers provided feedback on clicker class sessions.

All statistical analyses were performed using SPSS (IBM, Armonk, NY).

Figure 5. Percentage of collapsed COPUS instructor and student codes for clicker (\( n = 80 \)) and nonclicker (\( n = 184 \)) class sessions. Each horizontal row represents a class session observation. The instructor codes in (A) for clicker class sessions and (B) for nonclicker class sessions are organized by the collapsed code Instructor Presenting. The student codes (C) for clicker class sessions and (D) for nonclicker class sessions are organized by the collapsed code Student Receiving.
Because percent-code calculations can be difficult to interpret for individual codes (see Methods for further details), we also examined the percentage of 2-min time intervals that included the traditional instructional codes such as instructor lecturing (Lec) and student listening (L). Nonclicker class sessions showed a broader range of percentage of 2-min time intervals that included instructor lecturing (Lec) and student listening (L), and higher median values for these two codes (Figure 6A). However, when comparing means (Figure 6B), there were no statistically significant differences between non-clicker and clicker class sessions. As a result, students in classes both with and without clickers spend a similar mean percentage of 2-min time intervals listening (L).

Taken together, these results indicate that University of Maine STEM classes that use clickers (typically characterized by larger enrollments) displayed a narrower range of Instructor Presenting and Student Receiving collapsed-code behaviors compared with nonclicker class sessions. We also observe that clicker and nonclicker class sessions have a similar mean percentage of 2-min time intervals allocated to instructor lecturing (Lec). As a result, students in classes both with and without clickers spend a similar mean percentage of 2-min time intervals listening (L).

**Documenting Variation in How Clickers Are Implemented at a Campus-wide Level**

Because faculty may implement clickers in various ways, possibly in alignment with entirely different pedagogical strategies, we examined classes that used clickers in order to identify some common instructional modes. We first separated out class sessions in which the instructors allowed peer discussion during clicker questions from those that did not (Figure 3, further description in Methods), and called this first mode Peer Discussion. The majority of the Peer Discussion class sessions used a combination of both individual and peer discussion during clicker questions. For the class sessions that did not allow peer discussion during clicker questions, we subdivided the class sessions into Individual Thinking (no peer discussion during the entire class period) and Alternative Collaboration (no peer discussion during the clicker questions but peer discussion during other group activities in the class period). Comparisons of the collapsed instructor and student codes between Peer Discussion, Individual Thinking, and Alternative Collaboration class sessions revealed a range of instructional behaviors (Figure 7), with the Alternative Collaboration class sessions showing the lowest abundance of Instructor Presenting and Student Receiving collapsed codes.

To determine whether the three different modes of clicker use impacted the percentage of time allocated to traditional instructional practices such as instructor lecturing (Lec) and student listening (L), we examined the percentage of 2-min time intervals that included these two codes. Peer Discussion, Individual Thinking, and Alternative Collaboration class sessions are all characterized by a range of percent 2-min time intervals that include instructor lecturing (Lec) and student listening (L), with the Individual Thinking class sessions showing the highest median values (Figure 8A). In addition, Individual Thinking class sessions had a significantly greater mean percentage of 2-min time intervals with instructor lecturing (Lec) and student listening (L) compared with the other two types of class sessions that use clickers (Figure 8B, one-way analysis of variance [ANOVA], Tukey’s post hoc test, p < 0.05, in both cases).

By definition, the Individual Thinking and Alternative Collaboration class sessions had no opportunities for clicker-mediated peer discussion, and the Peer Discussion class sessions contained at least one episode that included peer discussion (Figure 3). Because this definition does not account for possible variation in clicker use within the Peer Discussion class sessions, we also examined all 112 clicker episodes that occurred in the 44 Peer Discussion class sessions. Clicker episodes were classified into the following categories: Individual Thinking Only, Peer Discussion Only, and Individual Thinking and Peer Discussion Combined.
A. Peer Discussion Class Sessions

B. Individual Thinking Class Sessions

C. Alternative Collaboration Class Sessions

D. Peer Discussion Class Sessions

E. Individual Thinking Class Sessions

F. Alternative Collaboration Class Sessions

Figure 7. Percentage of collapsed instructor COPUS codes for (A) Peer Discussion (n = 44), (B) Individual Thinking (n = 22), and (C) Alternative Collaboration (n = 14) class session observations, organized by percent Instructor Presenting. Percentage of collapsed student COPUS codes for (D) Peer Discussion, (E) Individual Thinking, and (F) Alternative Collaboration class session observations organized by percent Student Receiving. Each horizontal bar represents a different class session observation.

(figure 2). In Peer Discussion class sessions, 75% of the clicker episodes included the opportunity for students to talk to one another (figure 9), with the most common practice including Individual Thinking and Peer Discussion Combined.

In Spring 2015, the teacher observers provided specific feedback for each class session they observed via an online survey, with a portion of this survey specifically focused on clicker use. In particular, the teachers were asked to provide information about how challenging the clicker questions were for the students based on the voting results (figure 10). The survey responses suggested that Peer Discussion class sessions more commonly included questions that were challenging for students when compared with Individual Thinking and Alternative Collaboration class sessions.

Taken together, these results indicate there are three different predominant modes of clicker use: Peer Discussion, Individual Thinking, and Alternative Collaboration. Among the three different modes of clicker use, the instructors in the Individual Thinking class sessions are spending significantly more time lecturing (Lec) and the students are spending more time listening (L). Alternative Collaboration class sessions tend to include lower abundance of Instructor Presenting and Student Receiving collapsed codes, largely due to the non-clicker group activities. However, the presence of these activities does not result in significant differences in percentage of 2-min time intervals allocated to instructor lecturing (Lec) and student listening (L) between Peer Discussion and Alternative Collaboration class sessions. In addition to providing an opportunity for peer interaction, clicker questions asked during the Peer Discussion class sessions tended to include questions that provided greater levels of challenge to students.

Examining How Individual Instructors Vary Their Implementation of Clicker Questions

Because instructors can vary their instructional style, we also explored the variation in how individual instructors used clickers. For this analysis, we focused on instructors who were observed teaching with at least two clicker episodes, regardless of the type of episode described in Figure 2. This analysis included 25 instructors from nine different departments. Nineteen of these instructors incorporated at least one opportunity for peer discussion (figure 11). However, nearly all of the instructors used Individual Thinking Only at some point in their instruction, and this strategy accounted for more than 50% of the episodes for 16 of the 25 instructors.
The results from our study appear to confirm the often articulated concern that adding clickers alone does not guarantee that instructors will spend more time overall on active-engagement, student-centered instruction. We suspect that part of the reason our data do not show dramatic differences between class sessions that use and do not use clickers is that both types of classes had a large range of collapsed-code instructional behaviors (Figure 5) and instructors teaching with clickers allocate a similar percentage of 2-min time intervals to instructor lecturing (Lec) and student listening (L) when compared with the instructors of nonclicker classes (Figure 6).

DISCUSSION

Here we discuss the first observation-based, multidisciplinary study of clicker implementation in STEM classes across a single campus. Observations of STEM classrooms revealed that nearly a third of class sessions used clickers (Figure 3). A comparison of class sessions with and without clickers showed that both types of classes had a large range of collapsed-code instructional behaviors (Figure 5) and instructors teaching with clickers allocate a similar percentage of 2-min time intervals to instructor lecturing (Lec) and student listening (L) when compared with the instructors of nonclicker classes.

In general, the clicker questions were:
- Challenging for students: the class vote was often split.
- Easy for students: the majority of the students answered correctly.
- Cannot determine: the instructor did not talk about the class voting results.

![Figure 8](image-url)

**Figure 8.** Comparisons of the COPUS codes instructor lecturing (Lec) and student listening (L) for Peer Discussion, Individual Thinking, and Alternative Collaboration class sessions. (A) Box-and-whisker plots show the median and variation for the three classroom types. The line in the middle of the box represents the median percentage of 2-min time intervals for the class sessions in each group. The top of the box represents the 75th percentile, and the bottom of the box represents the 25th percentile. The space in the box is called the interquartile range (IQR), and the whiskers represent the lowest and highest data points no more than 1.5 times the IQR above and below the box. Data points not included in the range of the whiskers are represented by an “X.” (B) Mean percentage of 2-min time intervals with instructor lecturing (Lec) and student listening (L) code among the three modes of clicker use. Asterisks indicate statistically significant differences, one-way ANOVA, Tukey’s post hoc test, $p < 0.05$. Bars indicate SE.

![Figure 9](image-url)

**Figure 9.** Distribution of clicker episodes for Peer Discussion class sessions.

![Figure 10](image-url)

**Figure 10.** Observers described how challenging clicker questions were for students based on the voting results shared with the class for the Spring 2015 observations ($n = 41$ class sessions). The results are shown for each of the three clicker class session modes: Peer Discussion ($n = 21$ class sessions), Individual Thinking ($n = 14$ class sessions), and Alternative Collaboration ($n = 6$ class sessions).
not use clickers is because we observed three distinct modes of clicker use: Peer Discussion, in which students had at least one opportunity to talk with one another during clicker questions; Individual Thinking, in which no peer discussion was observed; and Alternative Collaboration, in which students had time for discussion, but it was not paired with clicker questions (Figure 3). Our results indicate that for the Peer Discussion class sessions, the majority included clicker questions that combine both individual thinking and group discussion (Figure 3) and that the questions tended to be challenging for students to answer (Figure 10). Furthermore, instructors in the Individual Thinking class sessions spent a significantly larger percentage of 2-min time intervals lecturing (Lec) and the students spent a larger percentage of 2-min time intervals listening (L; Figure 8B). Finally, the Alternative Collaboration class sessions tended to include fewer Instructor Presenting and Student Receiving collapsed codes (Figure 7), but the presence of these activities did not result in significant differences between percentage of time allocated to instructor lecturing (Lec) and student listening (L) between Peer Discussion and Alternative Collaboration class sessions (Figure 8B).

Inconsistencies in How Peer Discussion Is Used with Clickers

Even though our data indicate that there are a variety of ways clicker questions are being used, Individual Thinking was the only behavior in 28% of the clicker class sessions we observed (Figure 2) and was an instructional strategy used by the majority of instructors (Figure 11). Although these instructors have successfully overcome many obstacles to the implementation of clickers in their classrooms, emphasis on the Individual Thinking strategy may inadvertently limit possible student learning opportunities. For example, if peer discussion is omitted, students may lose the opportunity to build scientific communication skills that are developed by articulating reasoning, evaluating the merits of others’ reasoning, and asking peers questions (Turpen and Finkelstein, 2010). In addition, performance increases attributed to peer discussion are lost (Smith et al., 2009; Lasry et al., 2009; Porter et al., 2011; Knight et al., 2015; Barth-Cohen et al., 2016). Furthermore, when peer discussion is omitted, faculty members do not have the opportunity to circulate around the class and listen to student reasoning (Mazur, 1997).

In addition to the lost student and instructor learning opportunities, previous work has shown that students have more negative attitudes about the utility of clickers when they do not discuss the multiple-choice questions (Keller et al., 2007). Student resistance often impacts instructional decisions, and faculty may abandon clicker-supported instruction and other active-learning pedagogies promoted by discipline-based education researchers (Silverthorn, 2006; Henderson and Dancy, 2007). It has also been documented that faculty prioritize personal experience over empirical evidence when making decisions about teaching strategies (Andrews and Lemons, 2015), and the negative experiences associated with nonoptimal clicker implementation may therefore have a long-lasting impact on future instructional decisions.

How Can We Encourage Faculty to Include Peer Discussion with Clicker Questions?

While research has shown there are benefits to allowing students to talk to one another during clicker questions (Smith et al., 2009, 2011; Porter et al., 2011; Knight et al., 2015;
Barth-Cohen et al., 2016), the observed variation in use of the peer discussion portion of clicker implementation is consistent with findings for other research-based pedagogies, which are typically changed and modified during implementation by faculty (Henderson and Dancy, 2007; NRC, 2013). Given that faculty often modify research-based instructional practices, how can we enhance professional development to make sure the peer discussion portion of clicker use is retained during implementation?

One response is to make sure faculty professional development motivates and targets the more nuanced aspects of effective clicker implementation. Considering that 86% of U.S. college faculty recently reported they were familiar with clickers (FTI Consulting, 2015), the majority of professional development audiences likely have some working knowledge of clickers. As such, it is important to move beyond dedicating an entire session to the basics of using a clicker system. Moreover, rather than polling professional development participants about whether or not they have used clickers before, the audience can instead be asked how they use clickers during instruction and can be asked to discuss with their neighbors and report to the group. This approach may reveal innovative ways in which people are using clickers, provides a more detailed picture of participant experience, and can serve as a launching point for motivating the value of peer discussion by drawing upon a combination of findings from research studies (Smith et al., 2009, 2011; Porter et al., 2011; Barth-Cohen et al., 2016) and engaging participants in activities designed to help them identify the features of clicker questions that encourage productive peer discussion.

In addition, recent work has shown that onetime faculty professional development workshops have a limited capacity to create change (Davidovitch and Soen, 2006; Henderson et al., 2011). Instead, faculty need ongoing, in-depth professional development and support (Henderson et al., 2011; PCAST, 2012). For this reason, at the University of Maine, we have started a yearlong faculty professional development program in which faculty meet in rotating pods of three: one individual teaches, one individual observes using the CO-PUS, and another provides feedback on areas identified in advance by the instructor. Notably, 89% of the faculty members participating in this program have said that encouraging student peer discussion is one of the predominant areas in which they would like assistance, and it will therefore be an ongoing focus of this program.

Finally, there is also a need for clicker question banks that are vetted by the community and include questions that have been shown to encourage productive peer discussion. The work described here indicates that instructors who are using peer discussion in their classes are asking questions that are challenging for students to answer (Figure 10), which are often time-consuming for instructors to write. In addition to presenting clicker questions, a question bank could also include aggregate student voting results, instructor reflections on how to most effectively follow up when student voting results are split among multiple answers, videos of students discussing the clicker questions with one another, and follow-up homework and exam questions that target the concepts from the clicker questions. These supplemental materials, in particular, could 1) foreground the ways in which clicker questions may be used to facilitate student learning, 2) provide some of the scaffolding needed to support effective implementation, and 3) serve as a flexible resource that faculty may adapt based on the needs of their classrooms.

CONCLUSION

Our campus-wide, observation-based study of clicker implementation in STEM classrooms revealed that instructors who used clickers demonstrated variation in implementation, with many instructors eliminating peer discussion during some if not all clicker questions. Omitting peer discussion impacts students' ability to articulate their reasoning and to work together to solve problems. In addition, instructors who omit peer discussion lose the chance to listen in on student reasoning and may encounter more student resistance to research-supported instructional techniques. To encourage faculty to include peer discussion, we recommend 1) focusing on peer discussion as an essential component of long-term clicker professional development programs that include multiple opportunities for faculty to learn about using peer discussion with clickers, and 2) establishing clicker question banks that include challenging, higher-order questions for faculty to adapt to their instructional needs.

ACKNOWLEDGMENTS

The authors appreciate the contributions of the middle and high school teachers who served as observers, including Elizabeth Baker, Toni Barboza, Tristan Bates, Michele Benoit, Stacy Boyle, Virginia Brackett, Elizabeth Connors, Tracy Deschaine, Lauren Driscoll, Andrew Ford, Bill Freudenberger, Kathryln Priest Glidden, Marshall Haas, Kate Hayes, Beth Haynes, Teri Jergenson, Danielle Johnson, Bob Kumpa, Lori LaCombe-Burby, Melissa Lewis, John Mannette, Nichole Martin, Lori Matthews, Nicole Novak, Jennifer Page, Patti Pelletier, Meredith Shelton, Beth Smyth-Handley, Damian Sorensen, Joanna Stevens, Rhonda Stevens, Nancy Stevick, Thomas White, and Michael Witick; and the University of Maine faculty members who welcomed the teachers into their classes for observations. We also thank Ken Akiha, Karen Pelletreau, and Mindi Summers for helpful feedback on this article. This material is based on work supported by the National Science Foundation under grants 1347577 and 0962805.

REFERENCES


better than staying quiet. AIP Conf Proc.


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 post-doctoral 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 adopt or adapt 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? Documenta-
tion 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:

the normal online submission process will be judged on the online dialogues, or the like. Meeting reports 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

---

**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.

**Announcements.** *LSE* accepts noncommercial announcements of meetings, workshops, and conferences and of funding opportunities and fellowships open to all.
Education guidelines. Prospective authors are advised that permission must be obtained in advance.

**Guidelines for Preparing Articles, Essays, and Features**

**General Instructions.** The following activities prior to submission of a manuscript to *LSE* do not constitute prior publication and do not preclude consideration of the manuscript by *LSE*: publication of a short abstract; presentation of data at a professional meeting or in a Webcast of such a meeting; or posting of a manuscript on an author’s personal website, in an online institutional repository, or on a freely accessible preprint server such as arXiv or bioRxiv. Publication of a paper in the proceedings of a scientific meeting generally does constitute prior publication. Authors should include copies of all closely related publications with their submission to *LSE*. A closely related publication is one that is in press or has been submitted elsewhere and includes some or all of the data presented in the manuscript submitted to *LSE*.

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 TIFF 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 supplementary material for further consideration by readers.

International authors may wish to consider using an editorial service, such as ScienceDocs (www.sciencedocs.com), EditBIOMED (www.editbiomed.com), The Medical Editor (www.themedicaleditor.com), American Journal Experts (www.journalExperts.com), Editage (www.editage.com), Bioscience Writers (www.biosciencewriter.com), Cognyte Ltd. (www.cognyte.co.uk), or Squirrel Scribe (www.squirrelscribe.com). *LSE* does not endorse any particular service, and cannot be held responsible for the services they provide.

**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 available (e.g., DVD, CD-ROM, PDF files, and html pages), how to request copies, and if there are any costs. *LSE* encourages provision of materials on a nonprofit
basis, but recognizes that this is not always feasible. Please contact LSE with any questions regarding this policy.

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 the 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 Book Reviews

Book reviews should include the complete book title, publisher information, author detail, ISBN, and retail price. The reviewer should suggest a title for the book review, but the title should be informative and descriptive of the book to most readers.

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.