HIGHLIGHTS OF 2012

Table of Contents

FEATURES

Editorial
Biology Education Research—A Cultural (R)evolution
   Erin L. Dolan ................................................................................................................................. 1–2

Letter to the Editor
Integrating Genomics Research throughout the Undergraduate Curriculum: A Collection of Inquiry-Based Genomics Lab Modules

Approaches to Biology Teaching and Learning
Common Origins of Diverse Misconceptions: Cognitive Principles and the Development of Biology Thinking
   John D. Coley and Kimberly D. Tanner ................................................................................................................................. 9–15

WWW.Life Sciences Education
Science Denial and the Science Classroom
   Dennis W. C. Liu ......................................................................................................................................................................... 16–21

From the National Academies
Changing and Evolving Relationships between Two- and Four-Year Colleges and Universities: They’re Not Your Parents’ Community Colleges Anymore
   Jay B. Labov .............................................................................................................................................................................. 22–29

Current Insights
Recent Research in Science Teaching and Learning
   Deborah Allen ............................................................................................................................................................................. 30–34

Book Review
Dynamic Classes and Eager Students
   Laura L. Carruth ......................................................................................................................................................................... 35–36

Meeting Report
GCAT-SEEKquence: Genome Consortium for Active Teaching of Undergraduates through Increased Faculty Access to Next-Generation Sequencing Data

Table of Contents Continued on Back Cover
Over 160 free online talks by scientists

- Martin Chalfie
  Green Fluorescent Protein

- Janet Iwasa
  Animating Cell Biology

- Matthew Meselson
  Replication of DNA

- A. Quiñones-Hinojosa
  How I Became a Scientist

- Ari Helenius
  Virus Entry

- Elizabeth Blackburn
  Telomeres and Telomerase

- Eva Nogales
  Electron Microscopy

- Roger Hanlon
  Camouflage in Cephalopods

Coming in 2013...a new site for all iBio talks, featuring new educational resources!

Supported by:

Find us on:
Apply to be a Biology Scholar!

The mission of the Biology Scholars Program is to empower biologists to be leaders in science education reform and catalyze professional societies to sustain undergraduate education reform.

Assessment Residency: Measuring Student Learning

Developed for individuals who are asking questions about the purpose of their courses AND the effectiveness of their course assessments. The goals of the Assessment Residency are to:

- Develop measurable learning outcomes for courses or course modules grounded in research on teaching and learning;
- Implement, integrate, and align assessment strategies with course learning outcomes;
- Explore opportunities to develop a research study analyzing student learning; and
- Create a community of practitioners to sustain improvements in teaching and learning.

Training Institute: June 12 - 15, 2013 | Washington, DC

Research Residency: The Scholarship of Teaching and Learning

Developed for individuals who are asking questions about the effectiveness of their teaching approaches. The goals of the Research Residency are to:

- Conduct evidence-based research on issues associated with student learning in biology;
- Identify collaborators from social, psychological, and statistical sciences education and anthropology;
- Make research results public through presentations and publications; and
- Create a community of practitioners that inquire about teaching and learning.

Training Institute: July 17 - 20, 2013 | Washington, DC

Transitions Residency: From Science Education Research to Publication

Developed for individuals who have initiated their research, collected preliminary data, conducted preliminary analysis of their data, and possibly presented their work to colleagues on campus, virtually, or at regional and national meetings. The goals of the Transitions Residency are to:

- Publish research on teaching and learning in biology;
- Identify collaborators for national and international STEM studies; and
- Create a community of practitioners that contribute to our understanding about teaching and learning.

Training Institute: June 24 - 26, 2013 | Washington, DC

Application Deadline for Residencies: February 1, 2013
Travel grants are available.

www.biologyscholars.org
biologyscholars@asmusa.org
This has been an exciting year in biology education research (BER). A National Research Council (NRC) report was published on discipline-based education research (DBER) that “investigates learning and teaching in a discipline from a perspective that reflects the discipline’s priorities, worldview, knowledge, and practices” (Singer et al., 2012). A review of the BER literature that was commissioned for the report identified 195 studies reporting data on student learning, performance, or attitudes in college biology courses (Dirks, 2011). Although relevant articles appeared in more than 100 different journals, most were published in just four journals: *Journal of Research in Science Teaching* (established in 1963), *Journal of College Science Teaching* (established in 1994), *Advances in Physiology Education* (established in 1996), and *CBE—Life Sciences Education* (LSE; established in 2002). LSE published an impressive 50 articles, 26% of the total, demonstrating that the journal has become a leading venue for publishing BER findings.

However, BER as a discipline is quite young (DeHaan, 2011). A number of recent reports have highlighted major issues in undergraduate biology teaching and learning that remain largely unexplored, such as the:

- consequences of teaching practices on the retention and long-term performance of biology majors (President's Council of Advisors on Science and Technology, 2012);
- differences in the motivations, experiences, and outcomes of different biology learners (e.g., majors vs. nonmajors, minority vs. majority students; Singer et al., 2012);
- mechanisms and outcomes of changes in teaching practices at the faculty and institutional levels (American Association for the Advancement of Science, 2011);

- development of biology majors’ computational, quantitative, and data-management and data-mining skills (NRC, 2003); and
- development of additional valid and reliable measures of biology student learning (Singer et al., 2012).

Generating this new knowledge will require expertise in biology, education theories, social science methodologies, and biology teaching that may only be realized through collaborations between biologists and social scientists. How can LSE help achieve this mingling of minds and perspectives? The journal will have to evolve as BER evolves. This evolution will not be easy, as LSE has two important sets of stakeholders: people who study biology teaching and learning and people who want to apply lessons learned from these studies in their teaching. Indeed, Singer and colleagues note (2012):

> Tension exists between publication venues that are intended to share research findings among researchers and venues that are intended to inform instructors of the findings of DBER that might be useful in their classrooms. Publications intended for practitioners to support change in classroom teaching generally earn less professional recognition than research-focused journals and may have lower standards for the rigor of
the research. High-quality research papers published in journals that practitioners are less likely to read may have less influence on classroom culture. (p. 37)

The editorial board of LSE has navigated this tension by publishing work that contributes to the development of knowledge and theory about biology learning, while also presenting practical strategies and outcomes that are approachable for biologists who teach. Indeed, results from a 2011 survey of LSE authors and readers revealed that authors (n = 227) chose to publish in the journal because it reached a biology researcher audience (89%), a BER audience (91%), and a biology teaching audience (90%). Readers (n = 664) indicated that they appreciated that the journal is written to reach a biology researcher audience (95%), a BER audience (92%), and a biology teaching audience (92%). Navigating this dualistic tension is growing more difficult, as the tools, methods, and results of BER become increasingly sophisticated. DBER scholars want more rigor and theory, while classroom practitioners want more real-world tools.

Perhaps the time has come for a cultural evolution, I dare-say a revolution, that prioritizes communication and collaboration between those who study biology teaching and learning and those who are positioned to apply the results of this work. This evolution will require us to establish new cultural norms. Biology education researchers need to embrace the roles of translator, ensuring that BER is meaningful and relevant to classroom practitioners, and critical friend, providing developmental feedback to biology researchers learning to study teaching and learning. The biology community needs to recognize BER as a legitimate career path, and to formalize opportunities for undergraduate and graduate researchers to pursue this path. The doctoral and postdoctoral training of biology researchers needs to include opportunities to become familiar with the BER theory, methods, and results that are important for good teaching. And reading BER literature needs to become an integral part of preparing to teach. As Editor, I will do my best to foster this coevolution by maintaining the original intent of LSE: to publish studies of biology teaching and learning that meet a high standard of quality, while remaining approachable to professionals engaged in biology teaching. The tent of LSE is big enough for both camps to share, learn, and grow together.

REFERENCES


Letter to the Editor

Integrating Genomics Research throughout the Undergraduate Curriculum: A Collection of Inquiry-Based Genomics Lab Modules

Lois M. Banta,*† Erica J. Crespi,‡§ Ross H. Nehm,† Jodi A. Schwarz,‡ Susan Singer,¶ Cathryn A. Manduca,† Eliot C. Bush, Elizabeth Collins,† Cara M. Constance,** Derek Dean,* David Esteban,† Sean Fox,§ John McDaris,§ Carol Ann Paul,† Ginny Quinan,† Kathleen M. Raley-Susman,† Marc L. Smith,‡ Christopher S. Wallace,§§ Ginger S. Withers,§§ and Lynn Caporale∥∥

*Department of Biology, Williams College, Williamstown, MA 01267; †Department of Biology, Vassar College, Poughkeepsie, NY 12604; ¶Department of Biology and Department of Human Ecology, Ohio State University, Columbus, OH 43210; ‡Department of Biology and Science Education Resource Center, Carleton College, Northfield, MN 55057; ‡Department of Biology, Harvey Mudd College, Claremont, CA 91711; **Department of Biology, Hiram College, Hiram, OH 44234; ‡‡Department of Biological Sciences, Wellesley College, Wellesley, MA 02481; ††Computer Science Department, Vassar College, Poughkeepsie, NY 12604; §§Department of Biology, Whitman College, Walla Walla, WA 99362; ∥∥Consultant; caporale@usa.net

We wish to let CBE—Life Sciences Education readers know about a portal to a set of curricular lab modules designed to integrate genomics and bioinformatics into commonly taught courses at all levels of the undergraduate curriculum. Through a multi-year, collaborative process, we developed, implemented, and peer-reviewed inquiry-based, integrated instructional units (I3Us) adaptable to a range of teaching settings, with a focus on both model and nonmodel systems. Each of the products is built on vetted design principles: 1) they have clear pedagogical objectives; 2) they are integrated with lessons taught in the lecture; 3) they are designed to integrate the learning of science content with learning about the process of science; and 4) they require student reflection and discussion (Figure 1; National Research Council [NRC], 2005). Eleven I3Us were designed and implemented as multi-week modules within the context of an existing biology course (e.g., microbiology, comparative anatomy, introduction to neuroscience), and three I3Us were incorporated into interdisciplinary biology/computer science classes. Our collection of

DOI: 10.1187/cbe.11-12-0105
§Present address: Washington State University, School of Biological Sciences, Pullman, WA 99164.
*These authors contributed equally to the writing of this letter.
Address correspondence to: Lois M. Banta (lbanta@williams.edu).
© 2012 L. M. Banta et al. CBE—Life Sciences Education © 2012 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.

Figure 1. Pedagogical elements of the I3U, which was based on the findings of America’s Lab Report (NRC, 2005) and was used as the primary curricular design framework for this project.
Comparison of a Highly Polymorphic Olfactory Receptor Gene Subfamily in Genetically Diverse Dog Breeds

Lois Banta, Norm Bell, and Duane Bailey
Willamgre College
Williamtown, MA 01267

Summary

In this three or four week project, students learn about single nucleotide polymorphisms (SNPs) by amplifying and generating sequence data on a highly polymorphic gene subfamhly in a diverse population of subjects (dogs) with which many students have considerable familiarity and affinity. In the first week, students make use of previously acquired knowledge of phylogenetic relationships and experience in sequence alignment to design primers specific for one subfamily of canine olfactory receptor genes. In the second week, each student uses his/her primers in a polymerase chain reaction (PCR) to amplify the corresponding DNA from one dog's cheek cells. During the following lab, PCR products are purified and the yield is confirmed on a gel. In the final week, commercial or in-house sequencing is used to determine the sequence of the PCR product. The data analysis draws on a published microsatellite genotype-based population structure of 85 domestic dog breeds, allowing the students to compare a phylogenetic tree estimated from a single gene with data obtained through a genome-wide analysis. An optional bioinformatics module introduces existing web resources to predict transmembrane domains and/or provides students with a short programming assignment in which they write a Perl script to perform this analysis on an olfactory receptor sequence.

Learning Goals

- Students will put into practice skills they have learned in previous bioinformatics labs, by analyzing and generating phylogenetic trees and by performing sequence alignments.
- Students will appreciate the difficulties in obtaining sequence data for an individual gene in a highly conserved family and the value of single-molecule sequencing strategies.
- Students will gain an understanding of the role of single nucleotide polymorphisms in gene family structure, the evolutionary processes that lead to genetic duplication and diversification, and the limitations in correlating function with specific SNPs.
- Students without prior wet-lab experience will gain familiarity with the rudiments of molecular biology tools and techniques in a relatively straightforward protocol.

Context for Use

This activity is part of an upper-level elective (12-14 students) in genomics and bioinformatics. It could also be used in a molecular evolution or a genetics course to integrate with classroom-based lessons on SNPs, PCR and/or DNA sequencing. The adaptability to larger classes is limited only by the availability of gel electrophoresis equipment and the cost of the PCR reagents and sequencing; students could also work in pairs to reduce expenses. Ideally, faculty would have access to as many different dogs as there are students or pairs of students; the larger the data set the more interesting the data analysis. However, multiple students or groups could also use the same dog's cheek cell sample. The lab presupposes no previous hands-on experience in molecular biology, and the only wet-lab manipulations (Weeks 2-3) involve setting up a PCR reaction, purifying the product, and running an agarose gel to confirm the yield. This lab was designed specifically to be a good first introduction to molecular biology, simple gel electrophoresis, and the use of pipettors for computer science, chemistry, physics and math students. The in silico investigation (week 3) builds on prior experience with Clustal and phylogenetic tree analysis, and the final data analysis reinforces the students' experience with Phylip or another tree estimation program (although Clustal can also be used to generate the neighbor-joining tree). The DNA sequencing can be performed in house, if the institution has its own sequencer, but the PCR products can also be sequenced commercially. The data analysis can be completed outside of lab.

Description and Teaching Materials

Week 1: Instructor introduces project, students design primers (computer lab needed)
- Behind the scenes between weeks 1 and 2: Instructor orders primers

Week 2: Instructor or students collect dog cheek cells from multiple dog donors and prepare genomic DNA (extract can be stored refrigerated for at least 1-2 days)
- Students set up PCR reactions

Figure 2. Excerpt from an activity sheet from the Genomics Instructional Units Minicollection describing one of the curricular modules developed within the Bringing Big Science to Small Colleges program (for the complete activity sheet, see http://serc.carleton.edu/genomics/units/19163.html).
<table>
<thead>
<tr>
<th>ID</th>
<th>I3U title</th>
<th>Conceptual content</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Reconstructing the Evolution of Cauliflower and Broccoli</td>
<td>Plant development</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Evolution</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bioinformatics</td>
</tr>
<tr>
<td>B</td>
<td>Human Single Nucleotide Polymorphism Determination</td>
<td>Genetics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Human evolution</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bioinformatics</td>
</tr>
<tr>
<td>C</td>
<td>Local Population Structure and Behavior of the Wood Frog <em>Rana sylvatica</em></td>
<td>Population genetics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Behavioral ecology</td>
</tr>
<tr>
<td>D</td>
<td>Comparison of Protein Sequences: BLAST Searching and Phylogenetic Tree Construction</td>
<td>Molecular biology</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Molecular evolution</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bioinformatics</td>
</tr>
<tr>
<td>E</td>
<td>Phylogenetic Analysis of Bony Fishes: Morphological and mtDNA Sequence Comparisons</td>
<td>Phylogenetics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vertebrate biology</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bioinformatics</td>
</tr>
<tr>
<td>F</td>
<td>Molecular Evolution of Gene Families</td>
<td>Genetics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Molecular evolution</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bioinformatics</td>
</tr>
<tr>
<td>G</td>
<td>Exploring the <em>Chamaecrista fasciculata</em> Gene Space</td>
<td>Plant genetics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Molecular evolution</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bioinformatics</td>
</tr>
<tr>
<td>H</td>
<td>Metagenomic Analysis of Winogradsky Columns</td>
<td>Microbial metabolism</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Community ecology</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ecosystems studies</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bioinformatics/programming</td>
</tr>
<tr>
<td>I</td>
<td>Behavior, Neuroanatomy, Genomics: What Can We Learn from Mouse Mutants?</td>
<td>Neurobiology</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Behavioral genetics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bioinformatics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Behavioral neuroscience</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bioinformatics</td>
</tr>
<tr>
<td>K</td>
<td>Comparison of a Highly Polymorphic Olfactory Receptor Gene Subfamily in Genetically Diverse Dog Breeds$^a$</td>
<td>Molecular evolution</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phyllogenetics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sensory biology</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bioinformatics/programming</td>
</tr>
<tr>
<td>L</td>
<td>Integrative Activities to Study the Evolution of Nervous System Function</td>
<td>Neurobiology</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Evolution</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bioinformatics</td>
</tr>
<tr>
<td>M</td>
<td>Modeling Molecular Evolution$^a$</td>
<td>Molecular evolution</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bioinformatics/programming</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Computer science</td>
</tr>
<tr>
<td>N</td>
<td>Constructing and using a PAM-Style Scoring Matrix$^a$</td>
<td>Molecular evolution</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bioinformatics/programming</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Computer science</td>
</tr>
</tbody>
</table>

$^a$I3U was implemented in an interdisciplinary biology/computer science course.
Table 2. Pedagogical attributes (scale of biological organization, genomic level of analysis, and bioinformatic skills taught) of I³Us developed in this project and disseminated on the project’s website

<table>
<thead>
<tr>
<th>I³Ua</th>
<th>Focal taxa</th>
<th>Questions asked at the level of</th>
<th>Analysis</th>
<th>Bioinformatics skills/tools</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Evolution</td>
<td>Behavior</td>
<td>Physiology</td>
</tr>
<tr>
<td>A</td>
<td>Brussica</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>B</td>
<td>Human</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>Wood frog</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>D</td>
<td>Fish/vertebrates</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>E</td>
<td>Fish/vertebrates</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>F</td>
<td>Xenopus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>Pea/variable</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>H</td>
<td>Eubacteria</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>I</td>
<td>Mouse</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>Worm</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>K</td>
<td>Dog</td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>Various</td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>Various</td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>N/A</td>
<td>√</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


Rapid advances in genome sequencing and analysis offer unparalleled opportunity and challenge for biology educators. More data are being generated than can be analyzed and contextualized in traditional teaching or research models. Indeed, this explosion of data has spawned rapid growth in the discipline of bioinformatics, which is focused on development of the computational tools and approaches for extracting biologically meaningful insights from genomic data. At the same time, access to vast quantities of genomic data stored in publicly available databases can offer educators ways to engage undergraduates in authentic research and to democratize research that was previously possible only at research-intensive universities with massive instrumentation infrastructures. The integration of genomic and bioinformatic approaches into undergraduate curricula represents one response to the national calls for biology teaching that is more quantitative and promotes deeper understanding of biological systems through interdisciplinary analyses (National Academy of Sciences, 2003; Association of American Medical Colleges and Howard Hughes Medical Institute [HHMI], 2009; NRC, 2009; American Association for the Advancement of Science, 2011). Yet relatively few faculty members who teach undergraduate biology have expertise in the fields of genomics or bioinformatics, and they may therefore feel inadequately prepared to develop their own new curricular modules capitalizing on this dispersed abundance of available resources.

Our Teagle Foundation–funded genomics education initiative, Bringing Big Science to Small Colleges: A Genomics Collaboration, was designed to address the challenges of helping
faculty members integrate genome-scale science into the undergraduate classroom. The goal of the project was to create and disseminate self-contained curricular units that stimulate students and faculty alike to think in new ways and at different scales of biological inquiry. To this end, a series of three workshops over 3 yr brought together a total of 34 faculty participants from 19 institutions and a diverse array of disciplines—including biology, computer science, and science education—to develop a set of lab modules containing a substantial genomics component. We believe that these modules are suitable for integration into existing courses in the biology curriculum and are adaptable to a variety of teaching settings.

The project website serves as a portal to activity sheets describing each I3U, complete with learning goals, teaching tips, and links to teaching materials, as well as downloadable resources and assessment tools (Figure 2), that can be customized by any interested educator. Each I3U was peer-reviewed by fellow participants, as well as by a professional project consultant who has extensive experience with Web-based description of teaching materials using this format (Manduca et al., 2006). The goals of this review process were to ensure that each I3U met the design criteria articulated above, and to evaluate whether the activity sheet provided both an easily accessible overview of the content and enough detailed information for other instructors to adapt and implement the material and its associated assessment strategies. This peer review was complemented by each participant’s own explicitly framed evaluation of his/her I3U through a formal reflection form (accessible at http://serc.carleton.edu/genomics/workshop9/index.html). Although these I3Us were designed for courses currently taught by the project participants within the specific institutions’ curricula, we propose that they can be inserted into other courses encompassing similar content (Tables 1 and 2) and/or learning goals (e.g., Figure 2). We have received many communications from colleagues at other institutions who have adapted our I3Us for their courses.

One fundamental characteristic of each I3U in our collection is the focus on guided inquiry. The benefits to an undergraduate of hands-on participation in research are well documented (Nagda et al., 1998; Gafney, 2001; Hunter et al., 2007; Kardash et al., 2008; Lopatto, 2009). Integrating authentic research experiences into the undergraduate curriculum allows this powerful learning model to be scaled from use with only a few students with use with entire laboratory sections (Lopatto 2009; Lopatto et al. 2008). Like other national participatory genomics teaching initiatives (Campbell et al., 2006, 2007; Ditty et al., 2010; Shaffer et al., 2010; HHMI, 2011), our model for curriculum development in genomics emphasizes synergies between student-centered research and education. However, in contrast with some of these other projects, our grassroots approach leveraged a wealth of existing expertise by providing opportunities for individual faculty members to develop, implement, modify, evaluate, and share undergraduate teaching modules that stem from their own research and/or teaching interests. In this regard, our project most closely resembles the Genome Consortium for Active Teaching, which provides faculty members and their undergraduates access to microarrays from a variety of organisms, allowing participants to define their own research questions in a model system with which they are already familiar (Campbell et al., 2006, 2007).

Our collaborative effort among biologists, computer scientists, and science educators has yielded a collection of pedagogical resources that can be adapted for use in a wide variety of educational settings. We invite other biologists to begin building on our work by using and providing feedback on our I3Us. Faculty who have tested materials that exemplify our design principles are encouraged to add them to our collection. For further information, please contact the corresponding author.

ACKNOWLEDGMENTS

This project was supported by a Teagle Foundation Fresh Thinking grant, along with supplemental support from Williams College, Vassar College, and Ismail Kola of Schering-Plough.

REFERENCES


L. M. Banta et al.


Feature
Approaches to Biology Teaching and Learning

Common Origins of Diverse Misconceptions: Cognitive Principles and the Development of Biology Thinking
John D. Coley* and Kimberly D. Tanner†

*Department of Psychology, Northeastern University, Boston, MA 02115-5000; †Department of Biology, San Francisco State University, San Francisco, CA 94132

Many ideas in the biological sciences seem especially difficult to understand, learn, and teach successfully. Our goal in this feature is to explore how these difficulties may stem not from the complexity or opacity of the concepts themselves, but from the fact that they may clash with informal, intuitive, and deeply held ways of understanding the world that have been studied for decades by psychologists.

We give a brief overview of the field of developmental cognitive psychology. Then, in each of the following sections, we present a number of common challenges faced by students in the biological sciences. These may be in the form of misconceptions, biases, or simply concepts that are difficult to learn and teach, and they occur at all levels of biological analysis (molecular, cellular, organismal, population, and ecosystem). We then introduce the notion of a cognitive construal and discuss specific examples of how these cognitive principles may explain what makes some misconceptions so alluring and some biological concepts so challenging for undergraduates. We will argue that seemingly unrelated misconceptions may have common origins in a single underlying cognitive construal. These ideas emerge from our own ongoing cross-disciplinary conversation, and we think that expanding this conversation to include other biological scientists and educators, as well as other cognitive scientists, could have significant utility in improving biology teaching and learning.

INTRODUCTION TO DEVELOPMENTAL COGNITIVE PSYCHOLOGY

Developmental cognitive psychology is a discipline that has not been closely connected to most biology education efforts, but may offer novel perspectives and insights. Cognitive psychologists study how organisms take in information about their environment, form internal representations of the information, and process or manipulate those representations to select and execute actions (Holyoak, 1999). Developmental cognitive psychologists study how these processes change over time as a result of environment, experience, and matura-
tion. The large and growing field that is developmental cognitive psychology includes several professional societies and a number of peer-reviewed journals (e.g., Journal of Cognition and Development, Cognitive Development) devoted to just this area of research.

Early research efforts in developmental cognitive psychology were led in the 1950s by Swiss researcher Jean Piaget, who was a psychologist and philosopher with early interests in zoology. While most scholars would agree that some aspects of Piagetian theory have become foundational assumptions underlying the entire field, other Piagetian ideas have largely been discarded. One aspect of Piagetian theory that is no longer widely accepted is the idea that development proceeds in qualitative stages, and that within each particular stage thought has unique qualities that apply across many different subject domains. Rather, modern researchers agree that development can be characterized via the notion of early partial competence (Smith et al., 1988). In other words, the
core competencies of many advanced cognitive abilities are present from early infancy, but nevertheless undergo substantial developmental change. One critical component of Piaget’s view that underlies virtually all modern research on cognitive development is the idea that cognitive development is an active process.

COGNITIVE CONSTRUALS—INFORMAL, INTUITIVE WAYS OF THINKING ABOUT THE WORLD

Many modern scholars believe that as children actively seek to understand, explain, and predict the world around them, they develop implicit or explicit informal theories about how the world works. As contrary evidence accumulates, children may or may not revise these theories. These theories give rise to what psychologists refer to as cognitive construals. A cognitive construal is an informal, intuitive way of thinking about the world. It might be a set of assumptions, a type of explanation, or a predisposition to a particular type of reasoning. Three such cognitive construals—teleological thinking, essentialist thinking, and anthropocentric thinking—may have particular relevance in understanding challenges and misconceptions commonly encountered in biology classrooms. In this paper, we attempt to make connections between each of these three cognitive construals and several areas of challenge in undergraduate biology teaching and learning. In addition, we explore how seemingly disparate biological misconceptions and misunderstandings may indeed have common origins in a single cognitive construal that undergraduate students may find implicitly useful in their thinking outside the realm of biology. In considering these explorations of the connections between biology education and developmental cognitive psychology, however, we suggest that readers keep in mind the following. First, there is not necessarily a simple one-to-one correspondence between a particular challenge or misconception and a particular cognitive construal. A given construal may give rise to a number of misconceptions and challenges, and any given misconception or challenge may be the result of multiple construals. Second, although all the cognitive construals that we discuss below are well documented in the cognitive developmental literature, modern research in cognitive development has focused largely on how these construals present in younger individuals, primarily during the period between birth and puberty. As such, in most cases little empirical attention has been paid to the developmental trajectory of these cognitive construals as students progress through middle school, high school, and college. In such cases, we will summarize what is known about the relevant arc of development in younger children, and extrapolate to our older populations of interest.

MISCONCEPTIONS RELATED TO TELEOLOGICAL THINKING

Consider which of the following statements you may have encountered in your own biology teaching and learning experiences. Some relate to molecular biology, others to transformations of matter and energy, and still others to evolution.

- Genes turn on so that the cell can develop properly.
- Birds have wings so they can fly.
- Plants give off oxygen, because animals need oxygen to survive.
- Individual organisms adapt and change to fit their environments.
- Evolution is the striving toward higher forms of life on earth.

These represent just a few examples of common biological misunderstandings encountered by teachers of biology from elementary school to college, as well as those promoting the understanding of science among the lay public. Biology instructors tend to perceive these challenges as unrelated and grapple with them in the classroom individually, as misconceptions in need of correction. However, these ideas may be more closely related in their origins than they initially appear. Specifically, all of these conceptual challenges relate to students’ need to answer the question of “Why?” in their studies of biology. Cognitive psychologists have shown that our minds are strongly biased toward causal explanations (e.g., Kahneman, 2012). We are quick to generate causal stories for any event brought to our attention: a slump by our favorite athlete, a perceived snub by a colleague, or a larger than average yield of zucchini from our garden. One common type of causal reasoning is known as teleological thinking, which is reasoning based on the assumption of a goal, purpose, or function. Kelemen (1999a) argues that teleological thinking is a central component of adults’ everyday thought. When reasoning about others’ behavior, adults make the teleological assumption that people’s actions are directed toward certain goals (e.g., he frequented the gym so that she would notice him; she saves money so that she can retire). Similarly, they presume that human artifacts, such as chairs and coats, are designed by their creators to fulfill some intended purpose (e.g., chairs are to sit in; coats are to keep us warm). As Kelemen emphasizes, teleological thinking provides an important component of adults’ intuitive interpretations of why events occur or why objects have the properties that they do.

The developmental arc of teleological thinking involves a pattern of “pruning.” Kelemen has shown that teleological thinking is widespread (or in her terms, “promiscuous”) among young children and becomes more selective and constrained with development (Kelemen, 1999b, 2012). For example, in one study of students in first grade through college, the youngest participants favored teleological explanations for a broad range of phenomena, including properties of nonliving objects (e.g., “The rocks were pointy so that animals wouldn’t sit on them and smash them”) and of animals (e.g., “Cryptoclidus had long necks so that they could grab at fish and feed on them”). College students were more selective; they rejected teleological explanations for nonliving objects, but 67–81% of them still preferred teleological explanations for biological properties (see also Kelemen and Rossett, 2009). Nor is this merely an educational by-product; Casler and Kelemen (2008) found that teleological explanations were common among Romani adults exposed to little or no formal education, suggesting that teleological thinking may be a basic feature of our cognitive architecture.

Teleological thinking is a widespread cognitive construal that is useful in helping us make sense of many aspects of the
world around us. However, this natural form of explanation is often extended inappropriately in the domain of biology. Students at all levels commonly explain biological structures and processes by reference to their supposed purpose, goal, or function. Indeed, this theme runs through the apparently disparate set of examples presented above. In all of these examples, biological phenomena are seen to be caused by the ultimate functions or outcomes of the phenomena. The first three examples occur on the cellular, organismal, and ecological scales, respectively, and all involve the use of an outcome as a causal explanation. The activation of genes results in proper development (ideally), but is caused by chemical signals and triggers in the cellular environment. Having wings certainly enables (some) birds to fly, but it is unlikely that selection forces acting on an avian ancestor anticipated this outcome and directed evolution accordingly. And students’ assertion that plants give off oxygen, because animals need oxygen to survive has been documented by multiple research groups (see Driver et al., 1994; American Association for the Advancement of Science [AAAS], 2012), impeding students’ ability to understand the biochemical origins of the oxygen released by plants, as well as the role of oxygen in cellular respiration within plants themselves.

The last two statements listed at the beginning of this section, seemingly unrelated to the first three, are common misconceptions about evolution that can also be linked to teleological thinking (e.g., Kelemen, 2012). The idea that organisms intentionally change their traits in order to better adapt to their environment and then pass these traits on to future generations is a well-known misunderstanding of what biologists refer to as adaptation (Bishop and Anderson, 1990; Passmore and Stewart, 2002; Stern and Roseman, 2004; AAAS, 2012). This is clearly teleological thinking; it substitutes a goal or purpose (better adaptation to the environment) for a causal explanation (variation and differential reproductive success). As such, it presents an intuitively attractive explanation for students, but can derail the development of their thinking about populations of living things, the differential survival of different members of a population, and how differential survival can drive changes in a population of organisms over time. Likewise, students who hold the common misconception of evolution as a collective striving toward “higher” forms of life—often including humans as the “most highly evolved” life form—mistake evolution for a purpose rather than a process.

In summary, teleological thinking may underlie a variety of seemingly unrelated biological misconceptions, and may thereby play a role in hindering students’ transitions from novice to more expert thinkers in the biological realm. No doubt many readers will be able to recall other misconceptions they have encountered in undergraduate biology education that appear to be associated with teleological thinking.

MISCONCEPTIONS RELATED TO ESSENTIALIST THINKING

Teleological thinking is probably not the only cognitive construal driving biological misconceptions. Consider the statements below, which do not appear to be mediated by teleological mind-set.

- Homeostasis keeps the body static and unchanging.
- Members of the same species are almost identical in their physical characteristics.
- If left alone, a wetland ecosystem will remain a wetland indefinitely.
- Because different cells in an organism have different physical characteristics, they must contain different DNA.
- Changing a single gene in an organism results in a new kind of organism.

The statements above represent misunderstandings that cross many biological domains, including molecular genetics, evolution, physiology, biotechnology, and ecology. What they share is the assumption that a core property or feature of a biological structure, species, or system determines its overt features and identity. This assumption may derive from a cognitive construal known as essentialist thinking. Essentialist thinking refers to a set of assumptions that people make about concepts. In cognitive science, the term concept is used in a more restricted way than in common parlance. For cognitive scientists, concepts are mental representations of categories, along with related knowledge. For example, our concept bird is a mental representation of everything we know about birds. Concepts can involve many different kinds of knowledge, including features (small, colorful, has feathers, flies), episodic memories (that flight of parrots you saw on Telegraph Hill in San Francisco), declarative knowledge (birds are dinosaurs), or mental images. Critically, concepts are summary representations. We do not file away every single bird-relevant thought, fact, or experience we have ever had. Rather, we selectively represent salient knowledge about birds and myriad other entities in terms of averages (Murphy, 2002).

In cognitive science, essentialist thinking is a cognitive construal in which concepts include not only summary representations of knowledge about category members, but also an assumption that there is some unobservable essential property (an “underlying reality” or “true nature”) common to members of a category that conveys identity and causes observable similarities among category members (Medin and Ortony, 1989; Ahn et al., 2001; Gelman, 2003). This is not a metaphysical claim about the structure of the world, but rather a psychological claim about assumptions implicit in people’s representations of some concepts. Nor is it a claim that people are explicitly essentialist; essentialist thinking may be implicit in the way they represent and use knowledge. Nor do people even need to know what the essential property might be; they need only behave as if there is one.

One consequence of essentialist thinking is that an entity’s category membership is ultimately based on the presence or lack of an essential property, rather than on superficial features. If a pigeon somehow survived a tragic accident in which it lost its feathers, wings, and beak, we would probably agree that it was still a pigeon. Conversely, changes to an essential property should result in changes to category membership. If our pigeon undergoes substantial genetic mutation, and thereby loses its beak, feathers, and wings, we may concede that it is no longer a pigeon (Keil, 1989; Rips, 1989).

Another consequence of essentialist thinking is the idea that members of a category are relatively uniform with respect to shared properties. For example, if you learn that one robin has
enzyme PX42, you might think it likely that all robins would have enzyme PX42. If the properties of category members are caused by an underlying essential property shared by all category members, then the essence should give rise to similar properties in all category members (Rips, 1975; Osherson et al., 1990; Coley and Vasilyeva, 2010). In essentialist thinking, variability among members of a category is disregarded as noise, resulting in concepts that simplify the bewildering array of information in the world into manageable chunks. Moreover, the assumption that category members share an underlying essence that is responsible for category membership and observed characteristics allows us to instantly and effortlessly make inferences about novel exemplars: this tiger will behave in this way, that turkey will taste that way. In sum, essentialist thinking allows us to explain and predict an otherwise incomprehensibly complex world.

There is substantial evidence suggesting that essentialist thinking is an early and pervasive cognitive bias (see Gelman, 2003). For example, by age two, children readily infer that living things from the same category will share internal features and nonvisible functions despite differing appearances (e.g., Gelman and Coley, 1990; Gelman and Markman, 1986; Coley, 2012). Indeed, children are often overzealous in their essentialist thinking; in studies using a “switched at birth” paradigm in which they are queried about whether a child will share biological and learned properties with birth parents or adoptive parents, preschoolers (like undergraduates) believe that a child will resemble its birth parents in biological characteristics (such as eye color), but also believe (unlike undergraduates) that that child will resemble its birth parents in learned characteristics as well (e.g., beliefs and preferences; Solomon et al., 1996; Taylor et al., 2009).

We propose that essentialist thinking may apply to our intuitive understanding of biological entities and systems, as well as species. If underlying essential properties cause external features, then the outward characteristics exhibited by members of any biologically relevant category—be it cells, species, or types of ecosystems—should be relatively uniform, static, and predictable. This construal unites the first three seemingly unrelated misconceptions listed above. First, the mistaken perception that homeostatic processes keep the internal environment of organisms constant and unchanging belies a lack of understanding of the temporal and spatial dynamics of living systems. The second statement represents students’ tendency to overgeneralize the characteristics of the members of a species. Students often do not recognize the extensive variation among individuals of the same species, which in turn creates an impediment to understanding the role of variation in the mechanisms of natural selection (Greene, 1990; Anderson et al., 2002; Shtulman, 2006). The third statement represents the application of this cognitive construal to ecology, and reflects the common misconception that the “natural state” of an ecosystem is static, rather than a succession of different communities.

Essentialist thinking may also underlie the last two statements at the beginning of this section, both of which concern students’ understanding of the relationship between DNA and physical traits at the level of cells or organisms. In these two examples, essentialist thinking could lead students to assume that a simple one-to-one correspondence exists between 1) essence (DNA) and 2) observable properties and identity (physical traits and biological classification). This has two clear implications. First, entities that display different observable properties (physical traits) should show corresponding differences in essential properties (DNA). Likewise, changes in essential properties (DNA) should imply changes in observable properties (physical traits), and by extension changes in identity (biological classification). When coupled with the observation that students may readily replace their intuitive notions of “essence” with a quasi-scientific notion of DNA (see Gelman and Rhodes, 2012), these extensions of essentialist thinking may lead to the last two misconceptions listed above. Students often assert that different cells in a multicellular organism contain different DNA (Hacking and Treagust, 1984; Banet and Ayuso, 2000; Lewis and Kattman, 2004; Lewis et al., 2000; Smith et al., 2008; Shi et al., 2010). While consistent with essentialist thinking, this conceptual stance reveals a lack of awareness of or appreciation for differential gene expression, a key mechanism by which different subsets of genes are expressed in different cells of an organism, resulting in dramatically different features (e.g., shape of a neuron vs. shape of a skin cell, different leaf shapes in different parts of the same plant). Relatedly, in the era of biotechnology and genetic modification, it is striking that not just students, but the general public, may view modification of a single gene in an organism (e.g., the addition of the spider silk gene to goat embryos) as a fundamental change in the underlying essence of the organism, which implies a new biological identity and classification.

In summary, essentialist thinking represents a second cognitive construal that has been well studied in developmental cognitive psychology that may provide an underlying explanation for a variety of seemingly unrelated biological misconceptions. No doubt educators have encountered many other misconceptions among undergraduate biology students that may be the results of essentialist thinking.

MISCONCEPTIONS RELATED TO ANTHROPOCENTRIC THINKING

Thus far, we have explored two cognitive construals, teleological thinking and essentialist thinking, and their connections to biological ideas. No doubt there are others involved. Consider the final set of statements below, which do not immediately appear to be mediated by either a teleological mind-set or an essentialist mind-set.

- Disturbance in ecosystems has no beneficial role.
- Cell death in an organism is unusual and pathological.
- Sexual reproduction always involves two organisms mating, and therefore plants cannot reproduce sexually.
- Plants suck up their food from the soil through their roots.
- The males of any species are usually bigger and stronger than the females.

As in the preceding sections, these apparently disparate misconceptions span biological subdisciplines but may, in fact, have common underlying origins. The cognitive construal that we propose may underlie these misconceptions is known as anthropocentric thinking; this is simply the tendency to reason about unfamiliar biological species or processes by analogy to humans. Analogical reasoning—trying to
understand an unfamiliar idea or situation by comparing it with something that is more well known—is a common strategy used across many domains of learning. In biology, human beings are a familiar and accessible biological kind and are therefore a very tempting source of knowledge that is often misapplied to nonhuman living things. For example, Inagaki and Hatano (1991) found that Japanese children as young as six used their knowledge of humans to make guesses about how relatively unfamiliar organisms (a rabbit, a grasshopper, or a tulip) would react in novel situations. Although the children used specific knowledge to rule out implausible inferences, they often overattributed human characteristics to nonhuman organisms that were similar to humans (i.e., mammals) in cases in which they lacked specific knowledge. Conversely, reasoning about nonhuman species by analogy to humans can also lead children to underattribute biological properties to species that are highly dissimilar to people. Carey (1985) found that prior to the age of 10, children showed a regular decline in the attribution of biologically necessary properties to organisms that roughly corresponded to the organisms’ phylogenetic distance from humans. For instance, children acknowledged that people reproduce, eat, and have a heart, but were less likely to say that insects or worms shared these properties (see also Coley, 1995). Thus, up to about age 10, children decide whether an organism possesses a certain property based on their knowledge of whether humans have that property and the perceived similarity of the organism to humans.

Anthropocentric thinking has been shown to vary with experience and with cultural assumptions about the place of humans in the natural world. For example, Inagaki (1990) found that children who raised goldfish tended to reason about a similar but novel organism (a frog) by analogy to goldfish, whereas children who had no experience with goldfish did so by analogy to humans. Likewise, Ross and colleagues (2003) found that anthropocentric patterns of reasoning increased between ages six and 10 among urban children, decreased over the same age range among rural majority–culture children, and were nonexistent among rural Native American children. Together, these results suggest that anthropocentric thinking is a common cognitive construal used when faced with a lack of more specific biological knowledge. And as with our other cognitive construals, developmental psychologists have paid little attention to the development of anthropocentric thinking past the age of 10.

Viewing the misconceptions listed at the beginning of this section through the lens of anthropocentric thinking, things like “disturbance” and “death” are negative for humans, and therefore easily seen as inevitably harmful to a biological system or organism. However, biological research has demonstrated that these processes are key to the robustness of ecosystems and individual organisms alike. In the case of ecological disturbance, events like fires are critical for succession in many communities. In the case of cell death, there are particular genetic systems that program cell death events in the course of development to give rise to complex shapes like the fingers of a human hand. In addition, cell death is an everyday mechanism during the course of an organism’s lifetime, needed to rid the body of damaged cells that could become cancerous. In the very different biological arena of reproduction, sexual reproduction in humans and among other mammals involves two different organisms mating; by analogy, it is tempting to view mating as a necessary component of sexual reproduction. However, this generates an overly narrow definition of sexual reproduction, one that excludes the many examples among plant species that undergo self-fertilization (Driver et al., 1994). The idea that all organisms take in food in essentially the same way that we do—such as plants sucking up food through their roots from the soil—represents another inappropriate instance of anthropocentric thinking, one that creates a serious impediment to students understanding the diverse ways in which different organisms obtain food (see Driver et al., 1994; AAAS, 2012). The use of the term “eating” to describe any intake of water, gases, or other molecular components by any organism, as well as the misconception that plants “eat” sunlight and air, are both naïve ideas that belie the wonders of plant photosynthesis (the special cellular processes by which plants make their own food) and the critical role of plants as the source of carbon-based molecules for other organisms in many ecosystems (see Driver et al., 1994; AAAS, 2012). Finally, extrapolating the observation that male humans tend to be larger and stronger than female humans to other species is inappropriate. Not only are there numerous exceptions (e.g., insects, bats, rabbits, squires, hamsters, and whales), but this anthropocentric thinking could also lead to a fundamental misunderstanding of what it means to be male or female in different species of organisms (Driver et al., 1994).

In summary, anthropocentric thinking represents yet a third cognitive construal that may provide an underlying explanation for a variety of seemingly unrelated biological misconceptions.

**DISCIPLINARY BORDER-CROSSING AND ITS IMPLICATIONS FOR BIOLOGY EDUCATION**

In its more recent history, the field of biology education has focused much effort on identifying common misconceptions, uncovering biological ideas that are particularly challenging for students, and developing classroom-ready tools that can measure the presence, absence, or changes of these ideas, especially in the context of undergraduate biology classrooms. Much less attention has been given to developing (or borrowing) theoretical frameworks that might provide a more synthetic or unified set of hypotheses about why so many students seem to think the way they do (Henderson et al., 2011; Coley and Muratore, 2012). Some attempts have been made to explore the role of difficulties in thinking across size and scale as a potential unifying impediment in biological thinking, though this has been primarily employed with regard to misconceptions about energy and matter and the ideas of photosynthesis and cellular respiration (e.g., Wilson et al., 2006; Hartley et al., 2011). However, theoretical frameworks such as those embodied in the cognitive construals of developmental cognitive psychology—exemplified here as teleological, essentialist, and anthropocentric thinking—are largely invisible in the biology education literature (for exceptions, see Tamir and Zhar, 1991; Rosengren et al., 2012). Our purpose is not to propose that the cognitive construals presented here are necessarily the only or even the most useful theoretical frameworks from another discipline that could be used by biology educators. Although we do think that these construals hold great promise and have attempted to provide...
corresponding evidence, our greater purpose is to encourage biology educators to engage in disciplinary border-crossing. The theoretical frameworks of many disciplines—including the cognitive construals of developmental cognitive psychology presented here—hold great potential for revealing common origins of apparently disparate student challenges in learning biology. In fact, such theoretical frameworks could provide an entirely novel approach to biology education reform, one that moves away from attempting to correct an ever-growing list of biological misconceptions piecemeal and instead moves toward engaging students in a systematic re-examination of deeply held intuitive ways of knowing—ways that are useful in everyday reasoning outside the classroom but might represent a stubborn impediment to the development of expert thinking in biological science.

REFERENCES


Science Denial and the Science Classroom

Dennis W. C. Liu

Howard Hughes Medical Institute, Chevy Chase, MD 20815

Biology teachers are accustomed to engaging individuals who do not accept biological evolution. Denial of evolution ranges from ignorance of the evidence to outright denial or distortion of data. The list of science denial topics has grown alarmingly over the years to include: HIV as the cause of AIDS, exaggeration of the health and environmental risks of genetically modified organisms, existence of holes in the ozone layer, the rise in antibiotic resistance, health risks caused by cigarette smoking, exaggeration and denial of harmful side effects of pesticides, water and environmental damage caused by hydraulic fracturing, the fear that vaccines do more harm than good, and, of course, global warming and climate change. Teaching climate science has become so perilous in some school districts that the National Center for Science Education, long known for activism in the arena of evolution education, has greatly expanded efforts in the arena of climate (http://ncse.com/climate).

In the face of denial by a substantial portion of society, the natural tendency of a scientist or educator is to pile on the data and examples, deepen explanations, and reason rigorously. However, such strategies may not be the most effective and are certainly not sufficient. In this feature, I will explore the nature and roots of science denial and review resources that instructors and students can use to recognize denial strategies while deepening their understanding of the scientific process.

Filtering information is central to human nature, so there are issues of psychology, communication style, and the role of the media to consider. Clearly, no one is thoroughly rational and objective in all his thoughts and actions. Each of us is guilty of denying facts on occasion, perhaps when it comes to our pets, our health, or our favorite sports team. However, systematic and organized denial that employs strategies of persuasion to combat scientific evidence is another matter.

Denialism is the systematic rejection of empirical evidence to avoid undesirable facts or conclusions. Various observers have distilled a number of common strategies employed by science denialists. The Skeptical Science website (www.skepticalscience.com/5-characteristics-of-scientific-denialism.html), focused primarily on global warming, has a concise list of denialist tactics that I paraphrase here:

1. Conspiracy theories: Propose that a complex and secretive conspiracy accounts for an overwhelming body of scientific evidence and consensus. Question the quality of the science. Former South African President Thabo Mbeki based his denial of AIDS on conspiracy theories that appealed to legitimate historical perspectives.

2. Fake experts: Highlight views inconsistent with established knowledge, often complemented by denigration of established experts, including questioning credentials, integrity, and motives. The tobacco industry employed industry-sympathetic scientists to attack mainstream research.
3. Cherry-pick data: Draw on isolated papers that challenge the consensus view and neglect the broader body of research supported by hundreds of papers. A single paper (now discredited) suggesting a connection between vaccination and autism is a prime example that has contributed to a rise in measles in the United States and other countries due to fearful parents not vaccinating their children.

4. Impossible expectations for research: Set unrealistic standards for data that invalidate an entire body of research. Since empirical science generally works with probabilities, denialists exaggerate uncertainty and gaps in data. The tobacco industry and antievolutionists have employed this tactic.

5. Misrepresentation of opposition viewpoints: Employ logical fallacies, including an appeal to personal freedom or other core values. A common tactic is to associate an opponent with Nazi policy or even personal characteristics shared with Hitler or Stalin.

The appeal to personal freedom has been marvelously lampooned in the book *Thank You For Smoking* by Christopher Buckley. A clip from the film version (http://movieclips.com/XYW5-thank-you-for-smoking-movie-ice-cream-politics; Figure 1) is well worth 2 min of your viewing time. It includes the following critical dialogue between a lobbyist father and a young son:

Nick: Because I'm not after you. I'm after them.

Joey: But you still didn't convince me.

Nick: I didn't have to. I proved that you're wrong, and if you're wrong, I'm right.

Joey: But that's not what we're talking about.

Nick: Ah, but that's what I'm talking about.

Joey: But you still didn't convince me.

Nick: Because I'm not after you. I'm after them.

While emphasizing the value of choice, the dialogue also illustrates the tactic of simply changing the subject. Several websites delve into denialist strategies a little deeper. The Hoofnagle brothers (one is a lawyer and the other a physiologist) have been at the forefront of identifying and analyzing denialism. They manage a blog (http://scienceblogs.com/denialism) that includes a more detailed primer on denialist tactics (http://scienceblogs.com/denialism/about.php). You will note a tone in the Hoofnagles' blog that departs from journalistic or scientific detachment. Blogs such as Skeptical Science (which has an app for countering climate denial arguments at www.skepticalscience.com/Newcomers-Start-Here.html) and the Hoofnagles' Denialism are positioned in opposition to the harsh tone of denialist blogs, such as the Forces pro-smoking website (www.forces.org) and the Heartland Institute (http://heartland.org), which denies global warming.

Journalists and the mainstream media occupy an important position in mediating denialism messages. Despite the proliferation of blogs and social media, professional journalists have the ability to amplify and disseminate a message, and long-standing traditions of balanced reporting can work against true fairness and veracity. Denialists are well aware of the media ecosystem and construct rhetorical arguments to give the appearance of legitimate debate where none really exists. They appeal to journalistic practices to cover both sides of an issue, to include diverse perspectives, and to cover a controversy. The University of Wisconsin and the National Association of Science Writers have organized a meeting, Science Writing in the Age of Denial (http://sciencedenial.wisc.edu; Figure 2). The meeting will address the challenges journalists face with science denial, such as reporting in a politicized climate, journalistic ethics, false balance in reporting, persuasive writing, and covering controversy. Video recordings of the talks will be available from the website (at the time of this writing, the April meeting has not yet taken place).

The topic of climate change has become especially challenging for journalists. A number of websites have been developed to disseminate information to reporters and other highly engaged audiences. Climate Central (www.climatecentral.org) is a research and journalism organization established to provide clear and up-to-date information concerning climate and energy. The site offers high-quality content that includes news bulletins, interactive graphics, special reports, and videos. The online magazine *Grist* (http://grist.org) is devoted to independent green news and has a section devoted to climate and energy. American University's School of Communication provides information and effective communication on climate science on a website called Climate Shift (http://climateshiftproject.org). These websites devoted to better science communication allow students to see what good-faith argument looks like in contrast with denialism. For example, Chris Mooney, a blogger for *Discover* magazine (blogs.discovermagazine.com/intersection/2011/04/21/false-balance-in-matthew-nisbet-climate-shift-report) has accused Matthew Nisbet of Climate Shift of phony balance. These two are engaged in a healthy debate between individuals striving for genuine balanced reporting and differing on where to draw the line, irrespective of personal opinions and bias. A recent editorial published in the *Wall
Street Journal highlights the challenge of conveying to a lay audience the nature of scientific skepticism (http://online.wsj.com/article/SB10001424052970204301404577171531838421366.html). The editorial, signed by 16 individuals with various climate science credentials, employs nearly all of the tactics of denialism, including comparing the consensus view on global warming with Soviet-era Lysenkoism. At the time of this writing (March 2012), 2850 comments have been posted in response to the editorial. William Nordhaus responded to a misleading description of his own work in the editorial by writing “Why the Global Warming Skeptics Are Wrong” in the New York Review of Books, (www.nybooks.com/articles/archives/2012/mar/22/why-global-warming-skeptics-are-wrong). Denialists often misappropriate scientific skepticism. Climate scientist Richard A. Muller recently became a famous climate change skeptic (see www.scientificamerican.com/article.cfm?id=i-stick-to-science) when he was recruited to testify before Congress with expectations that he would support climate change denial. Muller had called Al Gore’s An Inconvenient Truth a pack of half-truths and had expressed concerns about the quality of temperature data used to compile warming trends. He stunned the House Committee when he testified that his independent temperature measurement research confirmed the data about which he had originally been concerned. Muller showed Congress the colors of a true scientist, skeptical and grappling with evidence, but never denying evidence, even if it is contrary to his personal leanings. You can read his testimony on the Scientific American website (www.scientificamerican.com/article.cfm?id=muller-hearing). Two other prominent global warming skeptics include Anthony Watts and his blog Watts Up With That (http://wattsupwiththat.com) and Steve McIntyre of the Climate Audit blog (http://climateaudit.org). Watts and McIntyre characterize themselves as skeptical on some climate change issues, and Muller agrees that they are skeptics not deniers. Unfortunately, the tone of some of their blog posts sound denialistic. Watts’s blog, for example, has a posting reacting to Nordhaus’s response to the Wall Street Journal editorial (wattsupwiththat.com/2012/03/03/why-william-d-nordhaus-is-wrong-about-global-warming-skeptics-being-wrong). It is important that students understand the difference between the essential skepticism that all good scientists need and being a denialist or a knee-jerk contrarian. The reactions to skeptics who part from denialist camps are so strong that the skeptics are often denigrated as apostates. Richard Cizik, president of the New Evangelical Partnership for the Common Good (http://newevangelicalpartnership.org/?q=node/6), dramatically resigned his post at the National Association of Evangelicals. The proximal cause for his departure was comments he had made on shifting views of same-sex marriage, but he had been embattled for years over his views on environmental stewardship and accepting evidence for global warming.

One of the best ways to learn how to distinguish between a denialist and an honest skeptic is to delve into a case history. We are fortunate to have a tour de force case study in the book Merchants of Doubt by Naomi Oreskes and Erik M. Conway (Figure 3). The authors explain “how a handful of scientists obscured the truth on issues from tobacco smoke to global warming.” The behavior and tactics of these scientists embody the worst aspects of denialism. This book documents the activities of individuals who were motivated by an ideological perspective that led them to betray core scientific values, believing their ends justified denialist means. Merchants of Doubt is an enlightening case study and a stellar example of rigorous scholarship in science and history. The companion website (www.merchantsofdoubt.org) is especially valuable for the list of Key Documents it provides, as well as reference websites under the Resources tab. Students...
Science Denial

The book Merchants of Doubt is a masterly work of science, history, and investigative reporting on the role of some scientists in denying global warming and other issues. The companion website has excellent primary and secondary resources relating to the authors’ voluminous research.

Figure 3. The book Merchants of Doubt is a masterly work of science, history, and investigative reporting on the role of some scientists in denying global warming and other issues. The companion website has excellent primary and secondary resources relating to the authors’ voluminous research.

The typical citizen is a victim, not a perpetrator, of denialist tactics. Scholars from multiple disciplines have been working to better understand how the public receives messages with scientific content. How does the average person understand uncertainty? How do temperament and cultural affinities affect perceptions? Anthony Leiserowitz, director of the Yale Project on Climate Change Communication, has produced an outstanding 5-min video Global Warming’s Six Americas (http://environment.yale.edu/profile/leiserowitz/multimedia/anthony-leiserowitz-on-global-warnings-six-americas). His research reveals that Americans fall into six categories when it comes to global warming: alarmed (18%), concerned (33%), cautious (19%), disengaged (12%), doubtful (11%), and dismissive (7%). He notes that the “alarmed” individuals are a highly engaged and motivated “issue public” and that nearly one in five people being alarmed is a significant proportion of the public. While the “dismissive” group is less than half as prevalent, they are also highly motivated and voluble. Encouragingly, people in all of the categories support changes to energy policy. What is the best way to shape global warming messages to reach the 75% in the middle? Leiserowitz was formerly with Decision Research, an independent research organization that studies how people perceive risk and make decisions. For example, Decision Research has a number of projects in the area of risk perception and communication (www.decisionresearch.org/research/risk), as well as applied research projects on the environment (www.decisionresearch.org/research/environment) and medical decision making (www.decisionresearch.org/research/medical).

A perceived conflict between religious beliefs and scientific information is the prime motivation for denying biological evolution. However, the motivations for denying global warming and vaccine usage are less obvious. One might think that everyone would be interested in honest assessments of various risks. Trying to understand why people want to deny particular findings of science has led to an entire area of study termed “cultural cognition.” In a nutshell, cultural cognition refers to the influence of group values on individual perception. Dan Kahan and Donald Braman are leading scholars in the Cultural Cognition Project (Figure 4). Their website (www.culturalcognition.net) includes a syllabus for undergraduate and graduate courses (www.culturalcognition.net/teaching). It is fascinating to browse the project pages (www.culturalcognition.net/projects), which include studies on attitudes toward HPV vaccination, nanotechnology, and gun regulation. A core finding of their research is that “citizens experience scientific debates as contests between warring cultural factions” (see www.culturalcognition.net/browse-papers/fixing-the-communications-failure.html).

Group affinities and cultural background are not the only factors that influence how people perceive risks and make decisions. There are fundamental aspects of human cognition that influence decision making. Consider this simple math problem: a bat and ball cost $1.10 in total; the bat is $1 more than the ball. What is the cost of the ball? If you casually and quickly thought “a dime,” you are in good company—over half of 3500 individuals sampled at eight prestigious universities gave that answer. If you paused first, and decided to fully engage your rational mind instead of using intuition, you easily solved the math and came up with the correct answer, a nickel ($1.05 + $0.05 = $1.10). The key
is getting people to slow down and activate their rational system, especially when facing a horde of overriding cues. The Harding Center for Risk Literacy at the Max Planck Institute for Human Development studies decision making with the hope of helping people learn how to make less biased decisions; their particular focus is on health issues. The Harding Center offers a quiz to assess your risk literacy (www.harding-center.com/what-you-should-know). Getting people to consciously think about their own attitude toward assessing risks is an important step toward rational decision making. Researchers agree that poor decision making is rampant, but they are divided over how to fix this problem. The Harding Center promotes a curriculum for teaching statistical thinking to young children (see www.nature.com/news/2009/091028/full/4611189a.html). To ensure that the best evidence is used to make healthcare decisions, the Cochrane Collaboration for healthcare (www.cochrane.org) provides easy access to the best evidence for good healthcare practice. Perhaps similar projects could be adopted for a variety of scientific topics to help better discern and emphasize scientific consensus.

In closing I would like to mention Ignaz Semmelweis, the nineteenth-century obstetrician from Budapest. Semmelweis began practicing obstetrics in Vienna’s free maternity wards in 1846. He was shocked at the high incidence of death among mothers following birth, recording 36 deaths among 208 mothers in his first month working at Ward 1. Upon investigation, he learned that the death rate at Ward 2 was much lower and that poor women preferred to give birth in the streets rather than go to Ward 1. A common explanation among nurses was the existence of a poisonous gas in Ward 1. Semmelweis suspected the high mortality to new mothers might have more to do with the fact that Ward 1 also had a morgue and that doctors performed both autopsies and births. Ward 2 had only midwives and performed only births. Semmelweis suspected the physicians were transferring something harmful from person to person. He ordered all medical staff to wash their hands with chlorinated lime water and for the ward to be scrubbed with calcium chloride. Within several months, he managed to reduce the death rate to a negligible level. The medical establishment of the day refused to accept Semmelweis’s published findings, despite the success of his methods at multiple sites. The tragic end of his life was influenced by his peers rejecting Semmelweis’ sound conclusions (www.historylearningsite.co.uk/ignaz_smemmelweis.htm). The Wikipedia article on Semmelweis is particularly good and includes a lot of data from Semmelweis’s studies and publications (en.wikipedia.org/wiki/Ignaz_Semmelweis).

In Semmelweis’s day, many doctors were offended at the suggestion that they might carry diseases and should wash their hands. Today, we are fortunate that we understand the germ theory and that medical practice is generally grounded in sound science. Nevertheless, some physicians practicing in the twenty-first century suffer from a perception deficit when it comes to hand washing. Steven Levitt and Stephen Dubner, known for their *Freakonomics* book, website, and radio programs, which explore economics from interesting angles, including the difference between people’s perception of their behavior and their true actions, presented information about physicians and hand washing; you can view a short video to appreciate their analysis (www.youtube.com/watch?v=AEkOmn5hJFU). In one
study, doctors reported washing their hands 73% of the time, while nurses observed the doctors washing only 9% of the time, a shocking disparity. *Freakonomics* reports on a successful campaign waged by Cedars-Sinai Medical Center in Los Angeles to improve hand-washing compliance by using humorous posters and screen savers throughout the hospital ([www.freakonomics.com/2012/01/24/how-to-get-doctors-to-wash-their-hands-visual-edition](http://www.freakonomics.com/2012/01/24/how-to-get-doctors-to-wash-their-hands-visual-edition)).

Hand washing and a perception deficit among physicians brings us full circle back to cultural cognition. The Joint Commission Center for Transforming Healthcare has found that the only sustainable way to enforce good hand-washing practice at hospitals is to instill a culture in which appropriate hygiene is the norm. Various campaigns, including frequent email reminders or posted signs, for example, can help but are not sufficient. Only constant reinforcement through interaction with peers results in lasting solutions ([www.centerfortransforminghealthcare.org/projects/projects.aspx](http://www.centerfortransforminghealthcare.org/projects/projects.aspx)). In case you were wondering, indeed, nurses comply better with proper hand-washing regimens than physicians ([www.nursingtimes.net/hand-hygiene-compliance-exploring-variations-in-practice-between-hospitals/1944149.article](http://www.nursingtimes.net/hand-hygiene-compliance-exploring-variations-in-practice-between-hospitals/1944149.article)).

In this feature, I began by considering organized and intentional denialism, about which every honest scientist and educator must be concerned. The denial of evidence strikes at the very heart of what science is and why we do research. In a politically charged atmosphere, it is important to step back and consider the many factors of cultural cognition and the psychology of decision making that influence denialism. It is encouraging to see the popularity of books such as Malcolm Gladwell’s *Blink* ([see www.gladwell.com/blink](http://www.gladwell.com/blink)) and Daniel Kahneman’s *Thinking Fast and Slow* ([see www.brainpickings.org/index.php/2011/10/26/thinking-fast-and-slow-daniel-kahneman](http://www.brainpickings.org/index.php/2011/10/26/thinking-fast-and-slow-daniel-kahneman)), which dissect our faulty decision-making faculties. It appears that the general public is interested in how the human brain works in order to improve personal and societal decisions. However, the consensus of expert opinion is that changing behavior and improving decision making is very difficult. One of the privileges of being an educator is that your students come with minds that are open to learning something new, such as an understanding of how science works and its role in society. Students could gain an appreciation of science by looking at some of the denialism materials presented in this review. You may be aware that there is a push to implement common education standards across the United States that is gaining ground state by state ([www.corestandards.org](http://www.corestandards.org)). Nonfiction science reading has a prominent place in the English language arts standards, which include emphasis on science literacy. Teachers and students who recognize the role of science in our society should be able to recognize a denialist tactic when they see it.
From the National Academies

Changing and Evolving Relationships between Two- and Four-Year Colleges and Universities: They’re Not Your Parents’ Community Colleges Anymore

Jay B. Labov

National Academy of Sciences and National Research Council, Washington, DC 20001

This paper describes a summit on Community Colleges in the Evolving STEM Education Landscape organized by a committee of the National Research Council (NRC) and the National Academy of Engineering (NAE) and held at the Carnegie Institution for Science on December 15, 2011. This summit followed a similar event organized by Dr. Jill Biden, spouse of the Vice President, and held at the White House in October 2010, which sought to bring national attention to the changing missions and purposes of community colleges in contemporary American society.1 The NRC/NAE event built on the White House summit, while focusing on the changing roles of community colleges in science, technology, engineering, and mathematics (STEM) education. An in-depth summary of the summit was prepared by the NRC and NAE for publication in late Spring 2012 by the National Academies Press (NRC and National Academy of Engineering, 2012). This paper provides a synopsis of that report, which is available at http://www.nap.edu/catalog.php?record_id=13399, and emphasizes how we can use the report to improve STEM education for our students, but also how much progress still needs to be made to realize this ideal.

As participants at the summit emphasized, the traditional notion of a cadre of students moving smoothly through a pipeline from high school to a community college and then to a 4-yr college or university is providing an increasingly inadequate and incomplete picture of today’s postsecondary students. The presentations from the summit, summarized here, collectively make a compelling case that it would be a mistake for any 4-yr institution to ignore or dismiss the importance of community colleges, with their many roles in and contributions to improving STEM education. Leveraging these new realities will create new opportunities for improving STEM education for a much larger and more diverse population of college students through strategic and dynamic partnerships between 2- and 4-yr colleges and universities. These findings have broad utility both for readers of CBE—Life Sciences Education in our efforts to improve biology education.
education in particular and STEM education more generally, and for the departments and institutions in which the strategies are employed.

BACKGROUND

Community colleges are becoming an increasingly important sector of the higher education community in the United States. With nearly 6.5 million students enrolled annually (46% of the nation’s undergraduates), the nearly 1200 community colleges across the United States educate large numbers of students at a much lower cost than 4-yr institutions (e.g., Center for College Affordability and Productivity, 2010a). Students at a community college are typically far more diverse than their 4-yr counterparts in the same geographic location. Forty-seven percent of all African Americans, 47% of all Asian or Pacific Islanders, 55% of all Hispanics, and 57% of all Native Americans who are enrolled as undergraduates in the United States are currently studying in the nation’s community colleges (compared with more than 3000 4-yr colleges and universities). Enrollments of these minority students are thus far higher in community colleges than in many non-minority-serving 4-yr institutions. Half of all baccalaureate degree recipients began their college careers at community colleges (National Commission on Community Colleges, 2008; Center for Community College Student Engagement, 2010).

Community colleges fulfill multiple missions. Some have established ongoing relationships with local community organizations, governments, and businesses, which allows them to respond quickly to changing community needs. They may retrain displaced workers in skills needed by local businesses and open gateways to individuals who would otherwise lack the preparation or financial resources needed to receive a college education (Boggs, 2010; Center for College Affordability and Productivity, 2010a, b). The College Board’s National Commission on Community Colleges (2008, p. 5) argues that community colleges “are the nation’s overlooked asset. As the United States confronts the challenges of globalization, 2-yr institutions are indispensable to the American future. They are the Ellis Island of American higher education, the crossroads at which K–12 education meets colleges and universities, and the institutions that give many students the tools to navigate the modern world.”

Community colleges are also playing increasingly important roles in preparing grade K–12 teachers. The community college system has long been involved with preservice preparation of teachers, including teachers of mathematics and science (National Science Foundation [NSF], 1998; Recruiting New Teachers, Inc., 2002; Townsend and Ignash, 2003; Barnett and San Felice, 2005, 2006; National Association of Community College Teacher Education Programs [NACCTEP], 2008a, b; Patton, 2008; Fathe and Kasabian, 2009). Nearly half of elementary and middle school pre-service teachers take some or all of their mathematics and science courses at 2-yr colleges (NSF, 1998, National Science Board, 2006). However, as is summarized in this article, multiple barriers remain, inhibiting successful transfer to 4-yr institutions for would-be teachers (see, e.g., Shkodrani, 2004).

Increasing numbers of community colleges also offer professional development for in-service teachers. According to the NACCTEP (2008a, b), many community colleges offer focused courses, workshops, and institutes that boost teacher competency, especially in math, science, and technology. And many 2-yr collaborative preservice and professional development programs between K–12 school districts and higher education involve community colleges.

The roles of community colleges in the preparation of students for the STEM workforce are also becoming more visible and essential. Nearly half of U.S. students with bachelor’s degrees in science and engineering attended community college at some point during their education (Tsapogas, 2004). Almost one-third of the recipients of science or engineering master’s degrees began their postsecondary education at a community college. As noted above, nearly half of the nation’s teachers, including teachers of science and mathematics, completed at least some of their mathematics or science courses at community colleges.

Some states now permit community colleges to offer baccalaureate degrees in certain fields (Floyd et al., 2005; Lewin, 2009; Russell, 2010). Institutions that offer these degrees have established the Comprehensive College Baccalaureate Association to represent them. A few community colleges award graduate degrees. And, according to Dr. George Boggs, former president and CEO of the American Association of Community Colleges, increasing numbers of students who have earned baccalaureate or advanced degrees are returning to community colleges to complete some aspect of technical or skills training (Boggs, 2010).

A significant reason for the increasing enrollment in community colleges is their lower cost compared with 4-yr institutions. Although the absolute numbers are changing rapidly due to the current reductions of support for higher education in local budgets, community colleges charge far lower tuition than their 4-yr public or private counterparts, sometimes by an order of magnitude ($2500 per year for community colleges vs. $7000 to $18500 per year for public universities [in-state and out-of-state students, respectively] and $26000 average per year for private universities).

This growth and transformation in the nation’s community colleges is allowing them many more opportunities to be partners in STEM education with 4-yr colleges and universities. However, many educators and education policy makers at local, state, and national levels are largely unaware of these changes. Differences in governance, financial support, institutional cultures, and the roles of faculty at the nation’s 2- and 4-yr postsecondary institutions have led to a series of challenging issues that can impede the successful recruitment of and completion of degree programs by students who opt (or are forced) to begin their college careers in community colleges. At the same time, resolution of these issues can significantly improve opportunities for STEM-oriented students. Some of these issues include:

- Interactions between 2- and 4-yr postsecondary institutions, including articulation agreements
- Interactions between community colleges and secondary education, including dual enrollments and crediting for high school students
- STEM education pathways and their effects on employment of community college graduates
• Mechanisms to aid community colleges in broadening participation for students, especially those from underrepresented populations
• Standards for community college faculty, including credentialing and the use of adjuncts
• Nature and quality of STEM instruction at community colleges (although this is an important issue at many 4-yr institutions as well)
• Availability and quality of student advising
• Quantification of the economic impact resulting from community colleges’ preparation of students for the workforce (at the community, state, national, and global levels)
• Quantification of the completion rate for students at community colleges, including whether the appropriate indicators for success are being measured
• Nature and levels of external funding for community colleges

Examining Important Issues Related to the Intersections and Increasing Interdependence of Community Colleges and 4-yr Colleges and Universities

Given the complexities of STEM higher education and the new roles that community colleges are playing, a coordinated set of targeted studies analyzing and synthesizing existing data, and efforts to communicate these findings broadly, would greatly benefit community colleges themselves, as well as the institutions and organizations with which they interact. This broader community consists of K–12 education; 4-yr colleges and universities; business and industry; and local, state, and federal governments and policy makers. With these perspectives in mind (coupled with earlier work from the NRC to address articulation pathways for students from traditionally underrepresented populations [National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, 2011] and for students in engineering programs [NAE and NRC, 2005]), the Academy approached the NSF to undertake one or more studies and related convening activities that would address the aforementioned issues as they relate to STEM education more broadly.

Following several discussions with program officers at NSF (Advanced Technological Education program and the Division on Undergraduate Education), the staff directors of the NRC’s Teacher Advisory Council, Board on Higher Education and Workforce, Board on Life Sciences, and the Education Program Office of the National Academy of Engineering worked together on this proposal. Our goal was to organize and convene a summit that would address the changing and evolving dynamics between 2- and 4-yr colleges and universities around STEM education and the opportunities that might be afforded to both sectors and their students and faculty through more strategic collaborations and articulation agreements.

That proposal was approved, and an organizing committee was assembled (see Acknowledgments for names and affiliations of committee members). The committee and staff chose three scholars to draft commissioned papers on 1) outreach and partnerships between 2- and 4-yr institutions; 2) transfer and articulation issues; and 3) developmental (formerly called remedial) courses, in this case with an emphasis on mathematics. These papers were shared with summit participants prior to the event. Those draft papers, along with the summit agenda, biosketches of organizing committee members and presenters, and additional resources, such as PowerPoint presentations and videos from the webcast of the plenary sessions, are all available on the summit’s website. Revised versions of the commissioned papers are included as appendices in the summary report of the summit.

Some 100 invitees from a broad spectrum of experts in K–12 and higher education and education research and policy; program officers from a number of departments and agencies of the federal and state governments; and representatives from private foundations, business and industry, and professional and disciplinary societies registered to attend the summit. At least 150 people were estimated to have participated via a live webcast of all plenary sessions.

A brief overview of the plenary presentations in the order in which they were delivered is provided below, along with observations from the oral presentations of break-out groups. (The Academy’s more detailed and expansive summary report weaves these sessions and participant observations into a narrative based upon a series of themes that emerged during the summit.)

STEM Education and the World of Work. In her presentation, “Community College Opportunities and Challenges in STEM,” Jane Oates, Assistant Secretary of the Employment Training Center of the U.S. Department of Labor, described the department’s interest in community colleges. The Department of Labor focuses its attention and programs on students who do not complete high school and adults who return to school for retraining after losing their jobs. Under the Trade Adjustment Assistance Community College and Career Training Grant program, the Department of Labor is awarding a total of $2 billion over 4 yr to help prepare students for successful careers in growing and emerging industries. The department has recently instituted rigorous evaluation of both grants and formula funds that are distributed to states. In the first round of these grants, consortia of institutions that formed around needs of particular sectors, such as advanced manufacturing, healthcare, and engineering, received about 60% of the $500 million distributed. These consortia developed new curricula based on the needs of employers and developed new methods of delivering educational content, such as online learning. Community colleges are the point of entry for many people who could benefit from such programs.

The Department of Labor has supported the development of other web-based tools as well, such as My Skills My Future, which allows people to see jobs that are currently available. Similarly, My Next Move allows dislocated workers, especially military veterans, to search by zip code for jobs and to match jobs with additional skills that they have

http://nationalacademies.org/tac.
http://nationalacademies.org/bhew.
http://nationalacademies.org/bls.
http://nae.edu/programs.aspx.
http://nas-sites.org/communitycollegeosummit.
www.myskillsmyfuture.org.
www.mynextmove.org.
acquired through postsecondary training of some kind. In collaboration with the U.S. Department of Education, the Workforce Innovation Fund\(^\text{10}\) showcases innovations and partnerships between the workforce sector and community colleges. "The time is right for us to talk about the rigor and the wonder and the innovation that are going on in community colleges," Oates said. "We have for too long seen them as a stepchild, and they can do amazing things."

**Challenges of STEM Education Pathways.** Eric Bettinger, Associate Professor for Education and Economics at Stanford University, discussed "To Be or Not to Be: Major Choices in Budding Scientists." His work involves analyses of students' choices of majors and how those decisions change over the course of a 2-yr and a 4-yr education (Bettinger, 2010). His data from Ohio are for the 1998–1999 cohort of incoming students who took the ACT precollege examination and indicated their initial preference for a major in college. The data set has enabled him and his colleagues to follow choices that these students made in subsequent years. In a total sample of 16,000 students, 8.0% indicated an interest in the biological sciences and 11.7% in the physical sciences and engineering. These numbers are 5.5% and 9.4% for those who attend community colleges. The students at 2-yr institutions have somewhat lower average ACT scores than the average for all students, but their aspirations and academic characteristics are similar to those in 4-yr colleges and universities.

Among all students who declared an intention to pursue a STEM major, only 43% were still in a STEM field at the time of their last enrollment, with the rest moving to other majors. However, only 14% of the students at community colleges who intended to major in a STEM field were still in a STEM field at the time of their last enrollment. Almost half of all students who left STEM switched to business majors (48.7%). Other popular majors for switchers included the social sciences (21.2%) and education (11.1%). Among 2-yr switchers, about 30% switched to business majors, and slightly less than a quarter each went to social science and education majors. Only 5.5% of non-STEM majors for all institutions, and only 3.4% of non-STEM majors at 2-yr institutions, ultimately declared a major in a STEM field. These statistics have critical implications for the United States, given the increasing demand for workers with qualifications and training in STEM fields. However, as Bettinger observed, some of the majors to which students switch, such as education, also have extensive course requirements, albeit they are not as sequential as those for STEM majors.

Bettinger's data also show that women are significantly less likely to stay in STEM fields, even among top students, which suggests, he said, that the culture of STEM may be a factor in their decisions. Since female students take STEM courses in high school and still express an interest in majoring in those subjects, the cultural problems would need to start or intensify in college for this explanation to hold. According to Bettinger's research, African American students in 4-yr colleges are less likely to defect from STEM majors than other students, especially among the top African American students. However, that is not true at 2-yr colleges, where there are no statistical differences between African American students and other students.

One factor in students' decisions about majors is the amount of money they potentially could earn after graduation. About three-quarters of all college students agree in surveys that an important objective of a college education is to be "well off financially" (Pryor et al., 2011), and some colleges have increased their focus on vocational offerings. This is especially observed at 2-yr colleges, as in most cases it is part of their mission.

**Three Primary Foci of the Summit.** The three authors of the commissioned papers, Becky Wai-Ling Packard (Mount Holyoke College), Debra Bragg (University of Illinois, Urbana–Champaign), and Alicia Dowd (University of Southern California), each briefly summarized their papers before engaging in a panel discussion with the audience.

**Factors Influencing Student Choices to Pursue STEM Degrees.** Becky Packard described an "ecological model" that examines the many environmental factors and the relationships among them that affect students' choices. Choices can be influenced by home, school, workplace, and other contexts, such as access to resources, transportation, financial aid, and child care. Many students, particularly first-generation and low-income students, do not know how to navigate the college and financial aid application process successfully. Financial considerations can be a significant barrier to college entrance and persistence. Further, students often lack information on transfer requirements and what they are likely to experience if they do transfer. When students gain mentoring in multiple contexts, they are more likely not only to persist in college but in a STEM major.

Packard highlighted several of the recommendations from her background paper. First, more students and families need to understand the differences between a technical degree from a career institute and the community college transfer pathway to a 4-yr STEM degree. They need to know much more about the broadening opportunities in STEM careers.
and should be exposed to STEM occupations; they will then be ready to learn what they need to do to qualify for those occupations. Excellent models already exist, and these are detailed in her paper.

Dual Enrollment. Any hands-on program designed to attract students into STEM needs to be paired with academic preparation, Packard said. She suggested an expansion in STEM-specific, dual-enrollment programs that pair community colleges or universities with high schools, in addition to more common outreach and recruitment strategies. Both honor students and struggling students can benefit from dual-enrollment courses, because taking college classes during high school can motivate students to continue their education. In addition, she contended that high school students should be able to use their college classes to fulfill a high school requirement, which would allow students greater flexibility.

Mentoring. Research on mentoring is robust, sophisticated, and rigorous, Packard noted. Most of the newer studies are comparative, longitudinal, or control for self-selection issues. However, more research is needed on how to create more effective mentoring programs and bring effective mentoring programs to scale. She noted that NSF has a mentoring requirement for grants that engage postdoctoral researchers, and suggested that there is no reason why this provision could not be extended to graduate and undergraduate students. Finally, informal mentoring and advising need to be infused by faculty into all courses, and especially at community colleges. Mentoring cannot be done through supplemental programs alone.

Packard noted that the American Institutes for Research has estimated that more than $4 billion in grants and state allocations are lost when new, full-time community college students do not return for a second year of study (Schneider, 2011). As a consequence, Packard stated, “The only thing more expensive than fixing retention in community college is not fixing it.”

The Challenges of Developmental Courses and Maintaining Student Success in Mathematics. In her overview, Debra Bragg said that many see mathematics as the backbone of the STEM pipeline. The typical mathematics sequence in U.S. education progresses from arithmetic to algebra to geometry to trigonometry to calculus. Many more students have recently embarked upon this progression in 2-yr institutions—from about one million students enrolled in 2-yr mathematics and statistics programs in the early 1980s to more than two million today. Over the past three decades, about 47% of mathematics enrollments in higher education have been at the 2-yr level. However, 57% of the students enrolled in 2-yr college mathematics are enrolled at the precollege, noncredit level. Elementary algebra, which is usually one to two levels below college-level algebra, commands the largest enrollments, and most students do not move beyond that level. About 7% of enrollments are in calculus, and another 7% are in statistics courses, with most students never moving beyond the introductory courses in these subjects. Bragg noted that relatively few 2-yr colleges offer special mathematics programs providing support for minorities or women (11% and 6%, respectively). In contrast, 90% of the 2-yr college mathematics programs require diagnostic or placement testing. An increasing number of researchers is raising questions about the use and value of these tests, said Bragg, and about their contributions to student defections from mathematics. However, about 14% of the mathematics programs offer undergraduate research opportunities, and 20% offer honors sections for their students. The American Mathematical Association of Two-Year Colleges has committed to reforming mathematics education to improve the overall situation.

Bragg made four suggestions for future action on the basis of her observations:

- Reform of the mathematics curriculum needs to encompass the entire educational system. Without a strategic, collaborative endeavor, it will be difficult for 2-yr colleges to implement and sustain reform, except in isolated ways.
- More research is needed on teaching and learning in 2-yr college mathematics, especially in college-level mathematics.
- The characteristics, experiences, and aspirations of students who enroll in 2-yr college mathematics need to be investigated further to understand how they develop the “habits of the mathematical mind” that are required to be successful in all STEM fields.
- Two-year faculty would benefit from opportunities to engage in research that encourages them to explore and assess new pedagogical strategies in the classroom and how these strategies affect student learning.

Challenges and Opportunities for Student Transfer from Community Colleges to 4-yr Institutions. In her analysis of transfer from community colleges to 4-yr institutions, Alicia Dowd cited a recent report from the National Science Board (2010) that called for 1) providing quality science and mathematics teaching to all students, 2) improving identification of STEM talent, and 3) creating supportive “ecosystems” through professional development for STEM educators. All three steps are needed to enhance the flow of students from community colleges to 4-yr institutions, she said. Unfortunately, current statistics are far from good news: The transfer rate for the most competitive private institutions has dropped from around 10% of student enrollments in 1990 to a little more than 5% in the most recently available data. Other institutions enroll a higher percentage of transfer students, but the percentages at these institutions have also been declining. Using survey data collected by NSF from recent college graduates, Dowd and her colleagues have examined degree choice among Latino and Latina students who earn an associate’s degree prior to transferring to a baccalaureate program. They found that the majority of students who transfer from a 2-yr college to a Hispanic-serving institution and earn a STEM degree do so in the social and behavioral sciences, rather than in engineering or the natural, agricultural, or environmental sciences. She reported that the culture, values, and beliefs of faculty are critical factors contributing to the lack of transfer students in the natural sciences and engineering. Faculty members from both 2- and 4-yr schools need to be partners in redesigning transfer systems, and they need robust evidence about what is effective and what is not. Transfer scholarships focused specifically on STEM fields could be powerful inducements both for students to pursue STEM degrees and for institutions to change their environments in
ways that would attract and retain these students. She also suggested that individual development accounts—savings accounts that are matched by public and private sources—could help increase the diversity of students in STEM fields.

Dowd also suggested the creation of evidence-based innovation consortia that would facilitate transformational educational innovations, thereby enabling all students to thrive. Consortia or networks composed of community colleges, universities, and open-education resource practitioners could support the adoption and adaptation of evidence-based innovations. These networks would include agencies, organizations, industry, foundations, and others interested in specific topics, such as the reinvention of the mathematics curriculum. They would support the development of effective tools for systemic interventions and could conduct and support research to gather and analyze evidence of innovations’ effects.

Developing Programs That Allow Underrepresented Minority Students to Thrive in STEM

As keynote speaker, Freeman Hrabowski, President of the University of Maryland, Baltimore County, began his remarks by noting that the demographics of the U.S. population are undergoing a dramatic shift. Ethnic groups currently underrepresented in the sciences soon will make up the majority of school-age children in this country. Drawing from the recent report of a committee that he chaired (National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, 2011), Hrabowski pointed out that the proportion of currently underrepresented minorities in science and engineering will need to triple to match their representation in the overall U.S. population. To maintain the strength and vitality of U.S. science and technology, many more of these minority children must not only decide to become scientists and engineers, but must be provided with opportunities and educational pathways that will allow them to succeed. Given the high representation of minority students in community colleges, these institutions will be critical in achieving this goal.

This underrepresentation of minorities in the science and engineering workforce stems from their low participation in science and engineering at every level of the pathways from elementary school to higher education and the workplace. Though underrepresented minorities now account for almost 40% of K–12 students in the United States, they earn only 27% of the associate’s degrees from community colleges, only 17% of the bachelor’s degrees, and only 6.6% of the doctorates in STEM fields.

In 2000, the United States ranked 20th in the world in the percentage of 24-yr-olds who had earned their first college degree in the STEM fields, Hrabowski noted. The report Rising Above the Gathering Storm (National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, 2007) called on the United States to raise the percentage of 24-yr-olds with a first degree in STEM from 6% to 10%. This would require a tripling, quadrupling, or quintupling of the percentages for underrepresented minorities, which are 2.7% for African Americans, 3.3% for Native Americans, and 2.2% for Latinos.

Since the 1980s, underrepresented minorities have aspired to major in science and engineering at about the same proportions as their white and Asian American peers, Hrabowski observed. Yet they complete STEM degrees in lower proportions than whites and Asian Americans. Five years after matriculating, only about 20% of underrepresented minorities who intended to earn a STEM degree have done so, compared with about 33% of whites and slightly more than 40% of Asian Americans. Hrabowski ascribed part of this attrition to the culture of science and engineering in college. A large part of the problem is the “weed-out” mentality still held by many college faculty in these subjects, he said.

The problem is urgent, Hrabowski said. A national effort to address underrepresented minority participation and success in STEM fields needs to be initiated and sustained. This effort must focus on all segments of the pathways, all stakeholders, and the potential of all programs, whether targeted at underrepresented minorities or at all students. Students who have had less exposure to STEM and to postsecondary education than others require more intensive efforts at each level to provide adequate preparation, financial support, mentoring, social integration, and professional development. Evaluations of STEM programs, along with increased research on the many dimensions of underrepresented minorities’ experiences, are needed to ensure that programs are well informed, well designed, and successful.

Colleges and universities need to increase the inclusiveness of their programs and the success of underrepresented students in STEM fields. College personnel have a tendency to say that the problem is at the K–12 level, but Hrabowski disagreed. K–12 education does need to be improved, but he contended that more students are better prepared to go to college than postsecondary faculty and administrators think. According to a recent study that he cited by Hurtado et al. (2010), the larger the number of Advanced Placement credits a student has taken, the higher her/his SAT score, and the more selective the university, the greater the probability that a student will leave science as an undergraduate. “It is not just a matter of preparation.”

Hrabowski cited several challenges that are particularly acute for community colleges:

- Inadequate levels of mathematical preparation. This is a problem for almost all colleges and universities, but it is an especially difficult problem at community colleges.
- Balancing the preparation of students for further study at 4-yr colleges and graduate schools, while also offering what for many students will be terminal 2-yr degrees and certificates for the technical workforce.

Several federal programs facilitate the transfer of underrepresented minorities from community colleges to 4-yr institutions, Hrabowski noted. These include programs such as the Bridges to the Baccalaureate and the Community College Summer Enrichment Program at NIH. Increasing numbers of community colleges have mounted promising initiatives, he noted, such as Miami Dade College’s Windows of Opportunity program, which helps academically promising, low-income students obtain associate’s degrees in the arts or in STEM disciplines; several programs at his own university encourage and facilitate student transfer from community colleges.

colleges. Strategies that promote transfer include grants that allow community college students to work less outside of their academic programs, enabling them to complete their associate’s degrees in 3 yr and then successfully transfer to complete their 4-yr degrees.

Hrabowski also emphasized the potential for internships to motivate students and prepare them for careers. Internships make students more serious about their work. The needs of industry can be infused into the curriculum, especially when people from business are involved in teaching the courses. Students learn how to work in teams, express themselves clearly, and gain other critical skills that they can use in the workplace (e.g., NRC, 2010).

CONCLUDING THOUGHTS

New realities in the postsecondary education landscape include the nation’s changing workforce needs and shifting economy and the increasingly pervasive roles that science and technology are playing in virtually all aspects of our society. These changes have created unprecedented challenges, but also unrealized opportunities to improve education and work opportunities for many more students, especially for those students who will pursue higher education at least partially through community colleges.

As noted in this paper, numerous obstacles are also limiting the realization of what is possible through enhanced cooperation between community colleges and 4-yr institutions. These barriers result in major disincentives for many students who are the nation’s untapped potential for the STEM workforce. In a presummit survey, several issues were articulated by registrants as requiring immediate attention, including a greater focus on inquiry-based learning in the classroom and lab; teaching STEM content in the context of employable skills; better articulation pathways between 2- and 4-yr institutions; providing better support systems for students and access to the scientific “culture”; making STEM education more visible to community college students and the potential of these students more visible to 4-yr institutions; and building professional communities across institutions to work on these challenges.

Although not always directed exclusively at STEM education, numerous projects and initiatives, supported by federal and state governments and private sources, are now addressing these issues. Through its Educate to Innovate initiative, the Obama administration has called for greatly increased funding of community colleges to produce more STEM graduates who can enter the workforce or pursue higher degrees in baccalaureate-granting schools. All but six states have now developed some level of articulation agreements that facilitate new opportunities to share resources and create new efficiencies in this era of shrinking budgets. Together, faculty and administrators from 2- and 4-yr institutions who are truly concerned about and dedicated to the improvement of student learning and academic achievement in the STEM disciplines will view one another as partners, collaborators, and colleagues.

This summit laid the groundwork for the further exploration of critical issues in this changing STEM education landscape. Although not included in the report, a postsummit survey of participants asking them to envision the next critical steps is posted on the summit’s website (http://nas-sites.org/communitycollege-summit). Hopefully, the combination of the summit report and the ideas that emerged from the survey will offer readers of CBE-LSE and their colleagues new avenues for discussion about these issues, which are of critical importance to postsecondary education today and in the future.

ACKNOWLEDGMENTS

The project on which this paper is based was supported by NSF grant EHR 1112988. I thank the members of the organizing committee for their continual support, ideas, and perspectives in organizing this event: George Boggs (Chair), American Association of Community Colleges (Ret.); Thomas R. Bailey, Columbia University; Lineea Fletcher, Austin Community College; Bridget Terry Long, Harvard University; Judy C. Miner, Foothill Community College; and Karl S. Pister, University of California.

I am grateful to my co-principal investigators for their ongoing support and hard work in organizing this event: Catherine Didion (National Academy of Engineering), Peter Henderson (NRC Board on Higher Education and Workforce), and Martin Storksdieck (NRC Board on Life Sciences). Margaret Hilton worked closely with me as the senior program officer and manager of the project. Cynthia Wei and Rebecca Fischler produced and managed the summit’s website. Orin Luke and Mary Ann Kasper served to organize the logistics for the summit.

I thank the presenters at the summit for their insights, which made this paper possible. I also thank Steve Olson for writing the first draft of the report from the summit. Some of the text in this paper is excerpted and modified from that report.

REFERENCES


12 www.whitehouse.gov/issues/education/educate-innovate.


Recruiting New Teachers, Inc. (2002). Tapping Potential: Community College Students and America’s Teacher Recruitment Challenge, Belmont, MA.


This feature is designed to point *CBE—Life Sciences Education* readers to current articles of interest in life sciences education as well as more general and noteworthy publications in education research. URLs are provided for the abstracts or full text of articles. For articles listed as “Abstract available,” full text may be accessible at the indicated URL for readers whose institutions subscribe to the corresponding journal. This themed issue focuses on recent studies about various aspects of authentic research activities experienced by undergraduate students and by primary and secondary school teachers. The authors discuss the implications of their findings with respect to the role of mentors, the participants’ perceptions about the scientific community of practice, and about the impact of mentored research on persistence and academic achievement in science.


The existence of well-established, federally funded education intervention programs for underrepresented minority students at the authors’ institution provided the opportunity to conduct a longitudinal study to examine the association between timing and extent of students’ participation in authentic research and their academic success and persistence as science majors (particularly in the biological sciences). The authors performed statistical analyses on transcript and admissions data collected and aggregated from 6834 freshmen who matriculated at their institution during a 4-yr period with a declared major in biology or in a related, biology-intensive major. The authors used the analyses to address the following research questions: 1) Is participation in undergraduate research (UR) positively associated with academic performance in biological sciences and with persistence to degree completion (of any degree, and specifically in biological sciences)? 2) If so, how do these associations compare when students from various underrepresented racial/ethnic minority groups are disaggregated from white and Asian students? 3) Does any association that is found vary with timing and duration of the research experiences?

For the analyses, the authors used logistic regression to estimate dichotomous dependent variables: graduation versus no graduation; graduation in biological sciences (vs. another degree program); and graduation in biology with a grade point average (GPA) of 3.0 or above (vs. below 3.0), which is a GPA considered to be a qualifier for graduate school admission. They used two major sets of independent variables: individual characteristics of the students (gender, race/ethnicity, socioeconomic status, indices of prior academic achievement) and research participation (timing and duration), with students’ GPAs in the introductory biology course sequence as an additional variable.

The major findings were that across all individual characteristics and differences in academic success prior to college, UR experience in this institutional context is positively associated with persistence to degree completion in any degree (including the biological sciences) and to academic success in biology. These findings persisted when the authors used conservative (highly restrictive) models for testing the associations. The increase in the probability of obtaining a degree between students who had research experiences and those who did not was largest for Hispanic and African-American students, indicating to the authors that research participation may be particularly helpful in preparing minority students for graduate study and careers in science. Additionally, the analyses revealed that participation in research more than once, or during or after the third year, was strongly associated with persistence toward degree completion; in some of the statistical analysis models, participation during the first 2 yr had either as strong or stronger associations. The analyses
did not detect interactions between racial/ethnic status and the timing and duration of the research experiences, a result that the authors attributed to the relatively small number of participants with these personal characteristics (i.e., a small cell size in the analysis).

Thus, in the context of this institution (a large, fairly selective research institution), the implicit assumption is that federally funded goals of programs aimed at providing research opportunities for underrepresented minorities (“to increase the number of underrepresented minority students who are credible candidates for post-graduate study,” p. 83) appear to be met. The authors interpret their results as providing evidence that making research experiences more available to undergraduates could serve to counteract the high attrition rates from science majors and to foster interest in science careers for more diverse student populations.


The authors examined the role that interactions between students and their research advisors play in acculturating UR apprentices to the community of science and shaping their scientist identities, as viewed through the lens of situated learning theory (Lave and Wenger, 1991). They interviewed 73 students (54% of whom were biosciences or bioengineering majors) participating in four UR programs from two research-intensive universities in geographically distinct locations, with the goal of informing the science community about how experienced researchers successfully foster the growth and development of undergraduates along the path toward becoming a scientist. Two of the sampled programs served large numbers of students from groups that are underrepresented in the sciences, resulting in a relatively diverse sample population in which 23% of the participants were African American and 12% were Hispanic (48% were women). Fifty-six percent of the study participants had completed at least three semesters and a summer of prior UR experience, and 44% of the interviewees were classified as being research novices (two semesters or fewer of UR experience).

The participants were interviewed once for a 40- to 80-min period; the authors used a semistructured protocol that allowed for follow-up probing and further exploration of participants’ comments. The interview questions focused on eliciting students’ perceptions of the value of participating in UR, their interactions with various laboratory personnel (research advisors, including graduate students and postdoctoral fellows, principal investigators, and other research group members), and the role that the experience was playing in shaping their identities as scientists.

Data analysis consisted of identification of coding themes under which the student comments fell (including a taxonomic analysis that clustered these into domains and subcategories), which the authors initially carried out using a qualitative software program, then discussed and refined. They used componential analysis to cluster comments for relevant group comparisons and also determined the frequency of student observations in each category.

An important finding from the study is that there were few identifiable significant differences between gender, race/ethnicity, and the type of UR program and institution with respect to students’ perceptions of the behaviors and practices of their advisors that best fostered their development as researchers. The variable that mattered, as revealed by intergroup statistical comparisons, was the extent of the interviewees’ prior UR experience. Novice students expressed needs that were different in major respects from those of the experienced students in all three of the domains articulated by the authors to categorize the practices by which advisors support undergraduate scientists-in-training. These three domains were professional socialization (transmission of norms and values, disciplinary knowledge, and skills), intellectual support (with problem solving and strategizing), and personal/emotional support (providing support by expressing interest and being friendly and accessible). In general, novices valued 1) clearly stated expectations and guidelines, 2) more extensive orientation to the concepts underlying their specific projects and the related language and tools, 3) understanding how the project fit into the “big picture,” and 4) getting past the intellectual frustrations and barriers of learning to think analytically and apply their knowledge to the specifics of the project. More experienced students expressed a greater need for support in areas related to their socialization, such as how scientific researchers think and act, on both a personal and professional level (e.g., how to deal with failures and setbacks and operate with intellectual independence).

The authors discuss the importance of the mentoring interactions for undergraduates from underrepresented groups, highlighting the strong potential of mentoring for building students’ confidence and interest in pursuing scientific careers. In discussing the implications of their findings with respect to informing use of successful practices, the authors also note the importance of increasing the awareness of graduate students and postdoctoral fellows (who commonly serve as UR mentors) of the educational role they also play as part of their everyday work as scientists.


[Full text available: www.ncbi.nlm.nih.gov/pmc/articles/PMC3284472/?tool=pubmed]

A model of organizational citizenship behavior (Organ and Ryan, 1995; McManus and Russell, 1997) and social exchange theory (Emerson, 1981) provided the conceptual framework for this study, which used analysis of data from a national faculty survey (DeAngelo et al., 2009) to predict individual characteristics and institutional contexts that influence faculty members’ decisions whether to sponsor authentic research experiences for undergraduates. These conceptual frameworks provided lenses for understanding why faculty may choose to include undergraduates in their research programs despite the many potential disincentives for doing so (including the way that institutional and departmental faculty reward systems are typically structured). The organizational citizenship lens would allow for the prediction that
a strong commitment to their institution and its mission, or a more positive perspective on undergraduates and the time they spend with them, for example, would motivate faculty to extend themselves beyond their official job responsibilities and obligations to sponsor undergraduate researchers. Social exchange theory would suggest such motivations as the belief that the relative benefits of playing this role in the professional development of young scholars outweigh the relatively high costs (in terms of the reward structure that influences how faculty distribute their workload time and allocate other resources).

The data source for the study was the University of California, Los Angeles, Higher Education Research Institute’s 2007–2008 Faculty Survey, which was administered to a national sample of faculty across institutional types and disciplines. The study also utilized an additional sampling of science, technology, engineering, and mathematics (STEM) faculty from institutions that confer relatively large numbers of STEM undergraduate degrees. The final analytical sample population consisted of 4832 STEM faculty survey respondents representing 194 colleges and universities. The authors used hierarchical generalized linear modeling (HGLM) as the main analytical technique, with a clustered design that nested faculty within institutions. The sole dependent variable for the analysis was the dichotomous issue of engagement versus nonengagement of undergraduates in faculty research projects. The authors grouped the study’s independent variables into eight blocks of faculty-level factors (predictors), as follows: 1) demographic characteristics of respondents; 2–3) aspects of faculty members’ careers (e.g., tenure status, rank, discipline, etc.); 4–5) various scholarly and teaching activities that might place constraints on faculty time (e.g., teaching interdisciplinary courses, collaboration with local community in teaching and/or research activities); 6) faculty research productivity; 7) indicators of faculty members’ goals for undergraduate education; and 8) faculty perceptions about the institutional climate for engaging undergraduates in their research. An additional block in the analysis consisted of institution-level measures, such as workload and mentorship activities, along with dichotomous measures that corresponded to different aspects of institutional type (e.g., private, doctoral/research university, or historically black college or university [HBCU], etc.). As part of the analysis, the investigators calculated factor scores in each of these blocks.

The HGLM analysis findings indicate that institutional context is strongly associated with faculty members’ likelihood of including undergraduates in their research programs, with faculty who teach at HBCUs, liberal arts colleges, and more selective institutions having the greatest likelihood. In particular, the results suggested a large gap (17 percentage points) between HBCUs and primarily white institutions with respect to the inclusion of undergraduates in faculty research. Faculty members who teach in the life sciences were more likely to include undergraduates in their research projects; for example, 20 percentage points more likely than faculty in the physical sciences, and 35 points more likely than faculty in the health sciences. Receipt of foundation or grant support and publication of a greater number of journal articles also had a strong positive association with decision to sponsor UR. While goals related to undergraduate education generally yielded mixed results, the factor that measured faculty members’ commitment to fostering scholarly habits of mind had a significant positive association with the outcome. Although the block of predictors related to faculty members’ perceptions of institutional climate also yielded mixed results, the results did suggest that faculty who expressed belief in three of the six categories (students at the institution were well-prepared academically, faculty at the institution are strongly interested in students’ academic problems, and departmental colleagues valued their research) were more likely to include undergraduates in their research programs.

The authors go on to discuss the implications of their main findings, including implications for increasing faculty commitment to the institution, faculty hiring decisions, and reshaping the faculty reward structure. They conclude by highlighting the importance of creating effective incentives to develop and sustain a viable UR program.


This study contributes to existing knowledge about the affective and cognitive outcomes of students’ participation in UR by exploring their preconceptions about the processes surrounding authentic science and how these beliefs might change over the course of a 10-wk research apprenticeship.

Study subjects were the 17 students selected from a 450-student applicant pool for participation in a federally funded, 10-wk summer research program with a chemistry focus at a public, research-intensive university. The sample population included roughly equal percentages of males and females of at least sophomore standing whose home institutional types varied; the students had no prior formal experiences with authentic research. Prior to the summer program, students were in electronic communication with their prospective mentors about the projects in which they would participate. At the end of a 1-d orientation program, students began interacting with their research groups, and after a short period of acclimation to their mentors’ laboratories, began work on their project within the first week.

Sources of data for the study include a modified version of a survey instrument developed to elicit beliefs about the nature of science (Nature of Scientific Knowledge [NOSK] survey; Rubba et al., 1981), periodic (at 2-wk intervals) structured interviews, directed journaling, and a follow-up questionnaire sent to participants several months following the summer research experience. The investigators administered the survey twice; the first administration informed interview protocols and assisted in the refinement of interview questions. The authors adjusted the interview questions to probe more deeply into students’ experiences and beliefs as their participation in the program progressed. The participants submitted the journals weekly, with entries consisting of responses to questions aimed at capturing beliefs related to such topics as the process of science and knowledge construction in science. The authors independently and iteratively analyzed the data sources for identification of emergent unifying themes, and reanalyzed the data after the coding themes were determined. They
cross-referenced across the various data sources to compile themes for individual participants in order to focus on the most significant changes in students’ beliefs that developed during the summer program.

To generalize broadly about the earliest preconceptions reported for the participants in this study: many thought that the authentic research would have major similarities to their laboratory courses—that it would be a well-planned and well-defined process governed by a single scientific method. Additional common preconceptions described by the authors include the notion that “everyday” research activity leads to groundbreaking results, and that it is a solitary activity guided by a quest for truth and understanding. Participants reported that the primary sources of these preconceptions were advisors, teachers, and friends who had participated in research projects, as well as textbooks and images from popular culture. Interviews conducted near the end of the summer research experience and responses to the follow-up questionnaire revealed that in many cases students considered their preconceptions to have been misconceptions and came to have what the authors considered to be slightly more mature understandings over the course of the 10-wk experience. Students came to appreciate the attention to detail and meticulousness needed in designing and conducting authentic research, particularly in the absence of the various fail-safe mechanisms of the standard laboratory course experience (laboratory manuals with step-by-step procedures, teaching assistants and professors as problem solvers), the relatively slow pace at which interpretable results emerged from the process, and the fact that successful research most often requires a team effort.

The authors conclude by discussing the implications of their results for how science laboratories are taught, suggesting that incorporation of strategies that more closely mirror authentic processes of science research might better shape students’ conceptions about how science is practiced. These results support the notion that initial preconceptions about how science progresses could play a key role in the design of effective learning experiences in the sciences, including UR programs.


[Full text available: http://ojs.jstem.org/index.php?journal=jSTEM&page=article&op=view&path%5B%5D=1523]

Like the research of Cartrette and Melroe-Lehrman summarized above, this study explored the nature of the initial expectations, motivations, beliefs, and attitudes undergraduates have about research internships and how these were altered by an actual internship experience in an interdisciplinary STEM field. Although the number of study participants was slightly greater (N = 25), and the research internship was of longer duration (an entire semester), the data sources used in this study were more limited in nature than the Cartrette and Melroe-Lehrman study, consisting solely of students’ guided journal entries made 3 wk after the start of the internship. Students were asked to reflect on their preconceptions and about whether their experience differed from these initial assumptions. The participating students, all of whom were enrolled at the institution that sponsored the research program, had a minimum GPA of 3.0 and at least a sophomore standing.

The authors performed a content analysis of the journal entries (initially, independently of one another), sorting and coding them according to themes that emerged from the data. In the study done by Cartrette and Melroe-Lehrman, several of the categories of preconceptions were related to beliefs about scientists, research environments, and the ease or difficulty of conducting research. That is, many of the participants initially viewed science as being a solitary endeavor, one conducted in a serious and stern environment that would require little use of communication and interpersonal relationship skills. Approximately half of the participants also thought that, like their course-embedded laboratory experiences, authentic science research projects would involve the following of set procedures, with mostly obstacle-free progress toward the expected results. As revealed in the journal entries, the actual internship experience enhanced these participants’ awareness that unlike the familiar terrain of laboratory courses, authentic research has creative, challenging aspects and can lead to unexpected results. Perhaps because of the timing of the reflection and reporting on preconceptions (3 wk into the experience), the authors were able to conclude that most of the participants’ preconceptions were contradicted by the actual internship experience.

The authors conclude by discussing implications from the study that could inform the faculty and staff who coordinate UR projects and programs. Largely, they highlight the value of incorporating an understanding of students’ preconceptions and expectations into conversations with students or other orientation activities prior to their internships, particularly with regard to their expectations about the nature of the supervision they will receive.

The following are recent CBE-LSE articles on UR:


Vol. 11, Highlights of 2012 33
I invite readers to suggest current themes or articles of interest in life science education, as well as influential papers published in the more distant past or in the broader field of education research, to be featured in *Current Insights*. Please send any suggestions to Deborah Allen (deallen@udel.edu).

REFERENCES


Feature
Book Review

Dynamic Classes and Eager Students


Reviewed by Laura L. Carruth, Neuroscience Institute, Georgia State University, Atlanta, GA 30303

Walking in to teach your first lecture as a newly minted assistant professor can be an intimidating experience. All eyes are on you, and the facial expressions of the students communicate emotions ranging from interest and excitement to boredom and sleepy disengagement. You know you want to hook your students on the class subject as soon as possible to encourage them to be more responsible for their own learning. One way to increase student engagement and motivation is to make use of a variety of techniques during the semester that encourage active learning. Elizabeth F. Barkley’s Student Engagement Techniques: A Handbook for College Faculty is a book that professors at all levels of experience in the classroom will find useful. The techniques described in this book are appropriate for all disciplines, on any campus, and Dr. Barkley does an excellent job utilizing examples from a variety of different classes and student ability levels.

Dr. Barkley’s intention is to encourage educators to move from the traditional mode (“lecturing” while “students listen,” followed by “testing”) to a more dynamic, learning-centered approach in the classroom. The goals of this handbook are to foster effective teaching in the college classroom by providing explanations of 50 effective learning activities.

The book is organized into three accessible sections—“A Conceptual Framework for Understanding Student Engagement,” followed by “Tips and Strategies,” and then “Student Engagement Techniques (SETS).” In part 1, Dr. Barkley introduces the reader to the various techniques by first defining student engagement and how engagement relates to academic success. She provides a conceptual framework for understanding and exploring student engagement and presents background information for teachers on the interaction of student motivation and active learning as a way to increase student engagement. She also provides a brief description of “What We Know from Neuroscience,” which some readers may find is written on a far too elementary level, especially in comparison with the sections of the text focused on ways of empowering students to be partners in their own learning process. Part 2 includes 50 tips and strategies presented to build on the framework laid out in the first chapters of the book, and will be especially useful, since many were derived from the science education literature.

DOI: 10.1187/cbe.12-01-0015
Address correspondence to: Laura L. Carruth (lcarruth@gsu.edu).

© 2012 L. L. Carruth. CBE—Life Sciences Education © 2012 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.
Several of these tips will have the most impact when reviewed before the semester begins and incorporated into course planning.

Part 3 of the book is a description of SETS. This section is devoted to describing 50 SETS, which are organized into two main categories: “Techniques to Engage Students in Learning Course-Related Knowledge and Skills” and “Techniques for Developing Learner Attitudes, Values, and Self-Awareness.” SETS within either of these areas are further divided into helpful groupings that allow the reader to quickly identify a possible SET for use in the classroom based on the needs of the instructor and students. Some SETS in category I emphasize content knowledge and are excellent for use as reviews or mini-assessments (e.g., the SET on Focused Reading Notes or Team Jeopardy), while others concentrate on critical-thinking skills, writing, problem solving, or performance. Category II SETS are designed to foster positive learning attitudes and values, self-awareness, and study skills in students. For all 50 SETS, essential characteristics are provided that include primary mode (i.e., individual or collaborative), activity focus (i.e., writing, reading, discussing, etc.), duration of activity (single or multiple sessions), and online transferability (low through high, with some SETS designed to be used online). In addition, each SET description includes a purpose, step-by-step instructions, examples from a variety of different disciplines, online implementation, variations and extensions, observations and advice, and key resources.

There are several SETS particularly appropriate for use in higher education science classes. These include using a Background Knowledge Probe to determine the level at which new information needs to be presented. The probes are instructor-generated questionnaires that can be used at the beginning of a course or at any point when new material is presented and are appropriate for use in introductory science classes all the way up to more advanced courses. Artifacts is another science-friendly SET that suggests using models and visual representations of information paired with guiding questions to stimulate student interest in a topic. SETS designed to address analysis and critical thinking include Classify, in which students use features of items (such as specimens or pictures of animals) presented in class to determine the classification and relatedness of the objects. Students can also identify component parts of an object (i.e., organs of an animal or organelles of a cell) and organize how the parts make the whole and influence classification. Others, like Team Jeopardy, are commonly used in classroom settings, but Dr. Barkley provides useful instructions that facilitate the success of this SET for student review. Team Concept Map is a SET that takes a commonly used educational tool, the concept map as a graphic organizer, and has students build it collaboratively. This is another SET that can be used in science classes of all types and levels. Several figures of sample team concept maps are provided to illustrate possible ways students can work together to organize information.

Six SETS are specifically designed to foster problem solving. Two examples include Think-Aloud-Pair-Problem Solving (TAPP), in which students partner to take turns problem solving out loud to address both the process and the product, and Send-a-Problem, in which students working in groups are given problems to solve, after which they then send both the problem and the solution on to the next group for analysis and evaluation. Of the 50 SETS described in this handbook, many examples can easily be adapted for a variety of student educational levels.

One area of the book, the section that addresses online implementation, would have been more helpful had it been expanded. In this section, Dr. Barkley does provide some information on how to convert a classroom SET into an online activity, but with new social media options available and more classes involving classroom technology, a greater number of concrete examples would have been appreciated. This handbook will be useful to educators with any level of experience in the classroom. It can provide guidance to new faculty and ways to refresh a stale course for faculty who want to update or enhance their classroom experiences. While some of the SETS are common sense—and you may be using them already—the help comes from the details about how to use the SETS more effectively and how to extend or vary different activities. Even if you do not have the time to appreciate the theory presented in the first sections of the book, the SET examples alone make this book valuable for novice and experienced educators alike.
To transform undergraduate biology education, faculty need to provide opportunities for students to engage in the process of science. The rise of research approaches using next-generation (NextGen) sequencing has been impressive, but incorporation of such approaches into the undergraduate curriculum remains a major challenge. In this paper, we report proceedings of a National Science Foundation–funded workshop held July 11–14, 2011, at Juniata College. The purpose of the workshop was to develop a regional research coordination network for undergraduate biology education (RCN/UBE). The network is collaborating with a genome-sequencing core facility located at Pennsylvania State University (University Park) to enable undergraduate students and faculty at small colleges to access state-of-the-art sequencing technology. We aim to create a database of references, protocols, and raw data related to NextGen sequencing, and to find innovative ways to reduce costs related to sequencing and bioinformatics analysis. It was agreed that our regional network for NextGen sequencing could operate more effectively if it were partnered with the Genome Consortium for Active Teaching (GCAT) as a new arm of that consortium, entitled GCAT-SEEK(quence). This step would also permit the approach to be replicated elsewhere.
engage students in the process of science and to present
science as a vibrant and active field (American Association for
the Advancement of Science, 2010). Extensive pedagogic re-
search concludes that participation in open-ended research
endeavors fosters a sense of ownership over a biological
subject, and enhances teaching and learning in biological
sciences (Teagle Foundation, 2007). Developing innovative
cross-disciplinary approaches and empowering faculty with
the tools to implement novel strategies remains a challenge
at all levels of undergraduate education.

In the last 5 yr, the rise of next-generation (NextGen) se-
quencing approaches in addressing biological problems has
been spectacular, but incorporating NextGen sequencing data
for active teaching in the undergraduate curriculum remains
a major challenge. For faculty at small and medium-sized
institutions of higher education, high teaching loads, lack
of access to state-of-the-art equipment, and budgetary con-
straints typically conspire to inhibit faculty from considering
NextGen sequencing in their own experiments. High capital
costs, extraordinarily high rates of technological change, and
dautning computational and analytical requirements make the
technology exceptionally challenging to assimilate into
the undergraduate curriculum.

The Genome Consortium for Active Teaching (GCAT; Camp-
bell et al., 2006) was developed a decade ago to meet
similar challenges in relation to the use of microarrays in
undergraduate biology education. GCAT offers highly dis-
counted microarray chips and array scanning and a support-
ing network of faculty expertise to educators working with
undergraduates. In one decade, the effort has trained over
360 faculty and 24,000 undergraduates in the use and
interpretation of microarray data. The newly trained students
were enrolled in primarily undergraduate institutions, in-
cluding those that historically serve underrepresented pop-
ulations. GCAT has met many of the goals of the BIO2010
report (Campbell et al., 2007), and recently expanded its
focus into synthetic biology (Wolyniak et al., 2010). Now
that NextGen sequencing is rapidly superseding microar-
ray technology for a variety of technical and economic con-
siderations, GCAT and others recognized the need to find
cost-effective and innovative strategies to facilitate active
teaching of NextGen technology at the undergraduate level.

Understanding the advantages and limitations of continually
evolving transformative technologies like NextGen sequenc-
ing is essential preparation for future life scientists, medical
professionals, and, indeed, a scientifically literate citizenry,
as the age of personalized medicine moves toward becoming
reality. In addition, analyzing raw sequence data provides
students with learning opportunities that underscore interco-
disciplinary concepts central and relevant to studies of all
forms of life.

THE MEETING

In this paper, we report proceedings of a workshop from a
National Science Foundation (NSF)–funded incubator grant
for research coordination networks for undergraduate biol-
ogy education (RCN/UBE). The network aimed to collab-
orate with a centrally located genome-sequencing core fa-
cility at the Pennsylvania State University (PSU) to enable
undergraduate students and faculty in the mid-Atlantic re-
region to access state-of-the-art sequencing technology. Initial
network participants included Juniata College, Susquehanna
University, Duquesne University, Hampton University,
Morgan State University, Ramapo College of New Jersey,
Gettysburg College, Lycoming College, Lock Haven Univer-
sity, Mount Aloysius College, Bucknell University, and Hood
College, with the genome-sequencing facility at PSU support-
ing the data acquisition and dissemination aspects of the ini-
tiative. The meeting was held July 11–14, 2011, at Juniata Col-
lege and PSU and included the individuals who helped write
the incubator grant and invited speakers Malcolm Campbell
(Davidson College), Anton Nekrutenko (PSU), Istvan Albert
(PSU), and Bill Morgan (College of Wooster). Through pre-
sentations and periodic whole-group discussions, we worked
together to exchange ideas to develop a structure to approach
the problem of introducing NextGen sequencing to under-
grades.

During the course of the meeting, a number of parallels
emerged between the thinking of the participants and the
philosophy of GCAT. Members of both groups valued the
academic freedom provided by their ability to choose and di-
rect their own research and scholarly activities. Both groups
recognized the value of communal support from colleagues
at similar small institutions to help compensate for lack of
a critical mass of peers on each campus. Both groups recog-
nized the need to partner with other groups, like the
microarray manufacturers in GCAT’s initial plan, or genome-
sequencing facilities, such as PSU. All of these considera-
tions suggested that the mid-Atlantic network for NextGen
sequencing could operate more effectively and enable the ap-
proach to be replicated elsewhere if it partnered with GCAT
as a new arm of that consortium. GCAT has established an ef-
ficient dissemination strategy through its website and listserv,
and many of the members currently using microarrays will be
poised to transition to NextGen sequencing as it replaces gene
chip technology. Our shared values and the success of the
RCN/UBE grant led to an agreement with Malcolm Camp-
bell for our RCN/UBE to become GCAT-SEEK(quence) and
to complement another GCAT initiative in the emerging field
of synthetic biology (GCAT-SynBio; Wolyniak et al., 2010).

At our network meeting, we formed a nascent community
of biologists from distinct areas (e.g., molecular biology, en-
vironmental science, plant biology, microbiology) aiming to
develop parallel research studies in the scholarship of teach-
ing and learning and discovery science. The specific goals of
the workshop were to: 1) learn lessons from the GCAT
model; 2) learn the scope of NextGen sequencing technol-
ogy, applications, and analysis; 3) develop common learning
goals for students using this technology and develop appro-
priate methods of assessment and; 4) develop goals and an
administration plan for the network.

Malcolm Campbell presented the keynote address on
GCAT, describing lessons learned from administration of the
consortium. In particular, he emphasized the importance of
an undergraduate focus, inclusion of minority-serving insti-
tutions, assessment of educational activities, advertisement
of the network, faculty development, and taking on the
most difficult problems related to a technology to make the
network valuable. The GCAT model also recognized the importance of the investigator retaining ownership in the direction of his or her research. This was one of the key reasons for adopting a model in which an investigator requested the raw sequence data related to their research expertise and passion. (For a detailed discussion of these considerations, see Boyle [2010].) The GCAT-SEEK network will periodically request proposals using its listserv.

The meeting included talks on NextGen technology and bioinformatics. Deb Grove and Craig Praul (codirectors of the PSU Genome Core Facility) detailed the latest sequencing technologies, applications, and costs associated with their Ion Torrent PGM, Roche 454FLX, and SOLiD 5500XL platforms. These are massively parallel DNA-sequencing machines capable of providing hundreds of millions to tens of billions of nucleotides of DNA sequence data in about a week. The resulting data must be processed using specialized bioinformatics techniques that may require high-powered computers. It was determined that up to 50% of costs could be cut when individual researchers cooperate to share sequencing runs. Unique DNA sequence adapters (bar codes) may be ligated onto an investigator’s DNA fragments before sequencing. This process allows each investigator’s samples to be individually labeled, pooled with other samples, and automatically separated after bulk sequencing. PSU Biochemistry and Molecular Biology faculty Anton Nekrutenko and Istvan Albert framed challenges and approaches to NextGen data analysis. Anton Nekrutenko described the Galaxy bioinformatics analysis framework that he and his colleagues developed (Blankenberg et al., 2010; Goecks et al., 2010). He emphasized the importance of the evolutionary underpinnings of bioinformatics analysis, and how comparison is key to understanding genomes. Istvan Albert, director of the Bioinformatics Consulting Center at PSU, suggested that bioinformatics analysis is challenging, because it is a highly interdisciplinary science incorporating information technology to manage data, computer science to analyze data, statistics to find meaningful patterns, and biology to form relevant hypotheses. He stressed that bioinformatics cannot be learned passively from a book but requires active-learning approaches and student commitment.

The challenge facing undergraduate faculty in introducing bioinformatics was addressed by Ash Stuart (Rampage College of New Jersey), Eric Sakk (Morgan State University), and Bill Morgan (College of Wooster), who is working on a related initiative in genomics education. The effectiveness of interdisciplinary approaches, open-ended inquiry, case studies, online student learning communities, undergraduate conferences, exchange of students between schools, and interaction among students in different disciplines on a single campus all had the potential to improve student communication, collaboration, and leadership skills. It was stressed that one of the desired skills students should acquire was adaptability, because the field of bioinformatics changes so rapidly. Having a forum to discuss the impact of genomic sciences and bioinformatics analysis using examples from the daily news was discussed as a way to connect the science to a broader societal context. Bill Morgan discussed his progress toward a free, online genomics textbook focused on interactive case studies with mathematical sidebars and modeled after the text, Discovering Genomics, Proteomics, and Bioinformatics, by Campbell and Heyer (2006). He has assembled a group of faculty members from throughout the United States with expertise in genomics to help coordinate the development of learning modules in 10 topic areas. The workshop participants were particularly interested in reviewing the learning objectives for the genome-sequencing topic. It was suggested that a template with learning objectives, protocols, and assessment instruments be developed for the next-generation sequencing module that would encourage faculty adoption because it would allow customization of its activities to the datasets of individual investigators.

No innovation in education can be considered successful if it is not subjected to rigorous evaluation and assessment in the context of defined learning objectives. The workshop participants discussed the learning objectives the network should have, and formed an assessment leadership team chaired by Tammy Tobin (Susquehanna University) and Jay Hosler (Juniata College). This team will guide development of appropriate instruments to monitor student outcomes for network participants, as well as to support individual faculty in assessing the impact of students working with raw sequence data in individual classes. Core learning outcomes proposed for the GCAT-SEEK network were the ability for instructors and students to do the following:

1. Explain each step in the generation and analysis of NextGen sequence data.
2. Discuss the basic biology assumptions that underlie sequence analysis (e.g., evolution, structure, and function).
3. Evaluate the strengths and weaknesses of the methods used in NextGen sequencing, including the impact that data quality has on bioinformatics analysis.
4. Construct a testable hypothesis and experimental design that uses NextGen sequencing and bioinformatics tools.
5. Choose and justify the appropriate methods for a specific NextGen sequencing application.

These proposed learning goals are also posted on the web (www.gcat-seeq.org). The goal is to have best teaching practices established and validated through appropriate assessment and distributed to all of the network participants and their colleagues.

**GCAT-SEEK: VISION AND APPROACH**

Following whole-group discussion, it was determined the agreed purpose of GCAT-SEEK is to 1) bring functional genomic methods into the undergraduate curriculum, primarily through independent and classroom-based student research using centralized core facilities to make NextGen sequence data accessible to undergraduates; 2) create a clearinghouse of information for educators to use when teaching NextGen sequencing and related topics; 3) create a large database of raw data and analyzed results for pedagogical use by GCAT-SEEK members; and 4) develop a global network of educators who are using functional genomics and NextGen sequencing in the undergraduate curriculum. GCAT-SEEK specifically aims to obtain group discounts at regional research-intensive core facilities, to negotiate software discounts, and to garner support for mini-grants to help cover the cost of the
initial sequencing runs for network participants. The network aims to support its members through online listservs, periodic workshops, and meetings, following an approach similar to that successfully used by other GCAT groups (Campbell et al., 2006). The network approach will add efficiency by coordinating projects and partnering investigators with appropriate sequencing platforms. Given that even the smallest purchasable unit of NextGen sequence will often contain a great excess of information for any given project and that many projects can be combined using bar codes (as discussed in the introductory paragraph), an organized staging for related samples from different investigators was envisioned. This may reduce cost of data acquisition to a few thousand dollars from departmental budgets or mini-grant programs. Furthermore, additional cost efficiency for network participants can be achieved through coordination of the synthesis and maintenance of a database of highly purified bar-coded primers for metagenomic analysis.

WHAT’S NEXT?

At a time of severe budget scrutiny, the efficiency and cost-effectiveness of the proposed approach is apparent. Sequencing cores are not running at capacity, and technological advances are lowering sequencing costs. Bioinformatics programs, databases, and computing requirements for many types of projects are all either already in the public domain or well within the budgets of even the smallest undergraduate colleges. Given sufficient interest, regional replication of some elements of GCAT-SEEK in the future should be considered as a means of lowering travel costs for meetings and workshops for participants and allowing students to more easily visit genome core facilities. Thus, with a modest investment, this program can start to meet the challenges of training the next generation of life scientists by engaging undergraduates in the process of science presented in the context of modern technology.

ACKNOWLEDGMENTS

We thank Anton Nekrutenko, Istvan Albert, and Bill Morgan for their excellent and insightful presentations. This meeting was supported by NSF Award DBI-1061893 to M.D.B.

REFERENCES


Essay

Writing-to-Learn in Undergraduate Science Education: A Community-Based, Conceptually Driven Approach

Julie A. Reynolds,* Christopher Thaiss,† Wendy Katkin,‡ and Robert J. Thompson, Jr.§

*Department of Biology, Duke University, Durham, NC 27708; †University Writing Program, University of California, Davis, Davis, CA 95616; ‡The Reinvention Center, University of Miami, Coral Gables, FL 33146; §Department of Psychology and Neuroscience, Duke University, Durham, NC 27708

Submitted August 1, 2011; Revised October 18, 2011; Accepted November 7, 2011
Monitoring Editor: Robin L. Wright

Despite substantial evidence that writing can be an effective tool to promote student learning and engagement, writing-to-learn (WTL) practices are still not widely implemented in science, technology, engineering, and mathematics (STEM) disciplines, particularly at research universities. Two major deterrents to progress are the lack of a community of science faculty committed to undertaking and applying the necessary pedagogical research, and the absence of a conceptual framework to systematically guide study designs and integrate findings. To address these issues, we undertook an initiative, supported by the National Science Foundation and sponsored by the Reinvention Center, to build a community of WTL/STEM educators who would undertake a heuristic review of the literature and formulate a conceptual framework. In addition to generating a searchable database of empirically validated and promising WTL practices, our work lays the foundation for multi-university empirical studies of the effectiveness of WTL practices in advancing student learning and engagement.

INTRODUCTION

A significant challenge in science education is how to move students from thinking about science as a collection of facts to be memorized toward a deeper understanding of concepts and scientific ways of thinking. Within undergraduate science, technology, engineering, and mathematics (STEM) education, one approach that has garnered considerable attention is learning-to-write—strategies designed to improve student scientific writing (Moskovitz and Kellogg, 2011). In contrast, there has been a relative neglect of writing-to-learn (WTL)—using writing to improve student understanding of content, concepts, and the scientific method. Despite substantial evidence that writing can be an effective tool in student learning and engagement (e.g., Poirrier, 1997; Bangert-Drowns et al., 2004; Brewster and Klump, 2004; Thaiss and Zawacki, 2006; Carter et al., 2007; Graham and Perin, 2007; National Survey of Student Engagement, 2008) and that WTL strategies can enhance knowledge acquisition and cognitive skill development in science disciplines (Rivard, 1994), WTL practices are still not widely implemented.

Rivard’s insightful review of WTL in science disciplines identified several key issues that impede widespread acceptance and application of research findings. Since different types of writing tasks result in different kinds of learning, we need to determine the links between writing and both critical thinking and conceptual change. Furthermore, writing practices need to be studied in context, rather than in isolation, and research designs need to examine the interactions among specific learning objectives, personal characteristics (e.g., prior knowledge), models of instruction (coverage vs. conceptual understanding), and specific writing tasks. The underlying metacognitive processes necessary for learning specific types of knowledge (declarative, procedural, and conditional) also must be identified and targeted by corresponding WTL strategies. Since higher-order thinking involves restructuring knowledge, we need to determine what types of writing activities evoke this process of knowledge...
transformation. Moreover, systematic, action-oriented research involving both qualitative and quantitative studies is needed to bridge the gap between researchers and practitioners. All these issues are still relevant today.

Given the promise of WTL and the specificity of Rivard’s recommendations for further research, what accounts for the lack of progress in the intervening 18 yr, and what new approaches will be needed going forward? We argue that two of the major deterrents to progress are the lack of a community of science faculty committed to undertaking and applying the necessary research, and the absence of a conceptual framework to systematically guide study designs and integrate findings. A third deterrent is the continuing disconnect between research and practice, which prevents instructors from identifying and incorporating appropriate WTL interventions. In an effort to address these issues, we undertook an initiative, supported by the National Science Foundation (NSF) and sponsored by the Reinvention Center (a consortium of 65 U.S. research universities dedicated to the improvement of undergraduate education at research universities), to build a community of WTL/STEM educators who would undertake a heuristic review of the literature and formulate a conceptual framework to guide collaborative studies and educational practices.

A COMMUNITY-BASED APPROACH

Although we acknowledge that some writing pedagogies can be resource-intensive to implement, there are ample sources highlighting more efficient and equally effective strategies for responding to student writing (e.g., Spear, 1987; Thaiss, 1998; Elbow and Belanoff, 1999; Ferris, 2003; Russell, 2005; Volz and Saterbak, 2009; Bean and Weimer, 2011). Therefore, we began with the premise that STEM faculty reluctance to incorporate writing in their courses derives largely from a lack of awareness of the research on the effectiveness of WTL, since most published findings are in journals not regularly read by STEM faculty and the majority of studies use methods unfamiliar to most scientists. Rather than simply reviewing the literature yet again and delivering “take-home messages” to STEM faculty (a traditional approach), we hypothesized that a more effective approach would be to engage STEM faculty directly in identifying promising WTL practices that improve undergraduate learning in STEM education (a community-based approach).

Our first step in building community was to form a WTL working group made up of 12 well-known experts in STEM research and education (Table 1). Its members formulated the intellectual framework for the project and conducted a heuristic review of the literature that had four specific objectives: 1) create a searchable database of WTL resources for both educators and researchers; 2) identify empirically validated and promising WTL practices; 3) determine critical gaps in current knowledge; and 4) lay the foundation for multi-university empirical studies of the effectiveness of WTL practices in improving student learning in STEM disciplines. The second step was to engage the STEM community in discussion of the most promising findings of the heuristic review and the implications for educational practice and research. Our method was to offer a workshop on WTL in STEM at the Reinvention Center 2010 Conference. The workshop was attended by 80 STEM faculty (the majority of whom were nominated by their universities’ Vice Provosts) who collectively considered how effective and promising WTL approaches could be applied in courses they teach, and who developed recommendations for the next steps in the research process to advance understanding of effective uses of WTL practices in STEM education. The postworkshop evaluation survey responses (n = 30) indicated that 76% of participants believe WTL will be an effective new tool in strengthening their students’ engagement; 90% expressed openness to experimenting with WTL practices and encouraging their colleagues to do so also; 79% expressed readiness to play a leadership role in the development of WTL at their institutions.

These findings attest to the value of the community-based approach. More specifically, by engaging the STEM community in both formulating the conceptual framework for the review of the literature and also in processing the findings with regard to the implications for both practice and subsequent research, we brought to bear a more diverse and inclusive perspective and yielded a set of recommendations more ready for implementation than the traditional approach of a single reviewer providing “take-home messages.” Furthermore, the community-building process resulted in faculty not only expressing their readiness to participate in the development and implementation of WTL practices on their campuses but also undertaking planning of multi-university collaborative initiatives.

CONCEPTUAL FRAMEWORK: CONNECTING WTL, NEUROCOGNITIVE DEVELOPMENT, LEARNING, AND TEACHING

Several key findings identified in the seminal National Research Council report How People Learn: Brain, Mind, Experience, and School (National Research Council, 2000) have implications for educational practices: Learning changes the physical structure and functional organization of the brain and people construct new knowledge and understanding based on what they already know and believe. These prior beliefs and knowledge can either facilitate or interfere with new learning. A related finding is that neurocognitive development continues through adolescence into adulthood, as the brain, particularly the prefrontal cortex, goes through a remodeling process; these changes in the brain are paralleled by changes in the cognitive abilities supported by these regions, particularly the development of cognitive skills involved in executive functions, social cognition, and self-regulation (Blakemore and Choudhury, 2006). This ongoing remodeling of the brain is the dynamic context in which undergraduate educational experiences are both impacted by and contribute to the development of higher-order cognitive processes and evaluative thinking.

Recent theory directs attention beyond the first-order cognitive processes that enable us to know about the world to the second-order metacognitive—“knowing about knowing”—processes that enable us to regulate cognitive, emotional, and motivational processes during learning (Kuhn, 1999). We now understand that successful learners are self-regulated, in that they employ a number of metacognitive processes while making meaning of information and their experiences. They elaborate on their existing knowledge, formulate relationships and
make connections among items, develop self-explanations, and monitor their own understanding and comprehension. There has been a corresponding paradigm shift in education from a focus on the curriculum and the acquisition of content knowledge to developing the learners’ metacognitive skills and learning strategies (Mayer, 1992) by incorporating modeling to make thinking visible and disciplinary practices overt, providing graduated supported practice (“scaffolding”), and encouraging reflection. Writing affords one of the most effective means for making thinking visible, and WTL practices can foster learning of both content and modes of thinking characteristic of disciplinary experts. These advances in understanding about how people learn provide the salient conceptual framework for a common—and compelling—research agenda that we propose take the following general form: What is the role of [specific WTL practice] in improving [disciplinary-specific learning objective] through impacting [specific cognitive, metacognitive, motivational, and/or emotional process], as a function of [context variables, such as course level and class size; discipline; level, background, and goals of students; and subdiscipline, local, and institutional factors]? Having a common conceptual framework for research enables STEM educators to undertake studies appropriate to their interests and particular context, while simultaneously participating in collaborative studies within and across universities, such that their findings contribute to the broader delineation and mapping of effective WTL practices.

LITERATURE REVIEW

Building on Rivard’s review, we focused our review on empirical studies published after 1994 in which writing strategies were designed to improve undergraduates’ learning in STEM disciplines. We examined 324 journal articles, books, book sections, conference proceedings, and reports that were identified through searches in the Web of Science and ERIC databases or suggested by the working group. Of these sources, 203 specifically focused on WTL pedagogies within STEM disciplines at the college level. We filtered studies through the lens of learning theory and used our conceptual framework to organize and categorize findings by level of course, discipline, and learning objectives. Representative studies reporting empirically validated practices, as well as descriptive studies that are promising and warrant further trials, were identified for each cell of the resulting matrix (Table 2). In addition, all studies were characterized by a number of additional key words to facilitate database searches (Table 3). The database is available at: http://bit.ly/fjudgo.

IMPLICATIONS FOR FUTURE RESEARCH DIRECTIONS

Our heuristic review found mostly descriptive case studies reporting on the effectiveness of particular WTL practices in improving students’ learning. Building upon emerging efforts supported by the literature to move the research toward the analytical and experimental levels, we offer the following recommendations.

First, the role of writing in improving learning needs to be reconceptualized. Learning is no longer understood as simply “acquisition of knowledge,” but as the construction of understanding and meaning as a result of social interaction. It is already well recognized that improving learning is no longer just a matter of strengthening associations and habits, but involves a change in understanding (Schoenfeld, 1999; Kuhn, 2005). The implications of writing assignments in STEM disciplines, therefore, should be reconceptualized to foster within students a shift from “knowledge telling” to “knowledge transforming” (Bereiter and Scardamalia, 1987). Second, to establish the links between writing and both conceptual change and critical thinking within specific disciplines, learning objectives need to be operationally defined in terms of the disciplinary content, conceptual knowledge, or the “ways of thinking” that characterize experts in the field and must include the underlying processes proposed to mediate and moderate the effect of particular WTL practices on student learning (Table 2). Although there has been an enduring focus in higher education on developing critical thinking and reasoning as general skills across academic disciplines, research is increasingly providing support for the view that reasoning is situation or domain specific (Beyer et al., 2007).

Third, studies must specifically seek to delineate the “mechanisms of effect” that is, the way in which a particular WTL practice brings about an improvement in student learning. How does writing “cause” learning to occur? Is it simply a matter of increasing time on task, or do students learn by applying cognitive and metacognitive strategies while writing? In addressing these questions, four interrelated systems have been demonstrated to affect learning...
Table 2. Key citations from the WTL in STEM bibliographic database, organized by learning outcomes, discipline, and course level, that represent exemplary descriptive studies, empirically validated studies, and promising practices.

<table>
<thead>
<tr>
<th>Content Knowledge</th>
<th>Biology/Life Sciences</th>
<th>Chemistry</th>
<th>Engineering</th>
<th>Math/Computer Science/Statistics</th>
<th>Physics/Earth Sciences</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Introductory</strong></td>
<td>Armstrong et al., 2008; Gerdeman et al., 2007; MacKay et al., 2005; Pelaez, 2002; Walvoord et al., 2008</td>
<td>Burke et al., 2006; Cooper, 1993; Margerum et al., 2007; Poock et al., 2007; Rosenthal, 1987; Shibley et al., 2001; Tilstra, 2001</td>
<td>Hanson and Williams, 2008</td>
<td>Ganguli, 1994</td>
<td></td>
</tr>
<tr>
<td><strong>Advanced</strong></td>
<td>Nekvasil, 1998; Ryan and Campa, 2000</td>
<td>Lillig, 2008; May et al., 2010; Stoller et al., 2005; Whelan and Zare, 2010</td>
<td>Nekvasil, 1998; Poronnik and Moni, 2006</td>
<td>Barr, 1995</td>
<td></td>
</tr>
<tr>
<td><strong>Capstone</strong></td>
<td>Schepmann and Hughes, 2006</td>
<td>Berry and Carlson, 2010; Ostheimer and White, 2005</td>
<td>Capstone: Codespoti, 1994</td>
<td>Capstone: Blakeslee, 1997</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Conceptual understanding</th>
<th>Biology/Life Sciences</th>
<th>Chemistry</th>
<th>Engineering</th>
<th>Math/Computer Science/Statistics</th>
<th>Physics/Earth Sciences</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Introductory</strong></td>
<td>Gerdeman et al., 2007; MacKay et al., 2005; Pelaez, 2002; Walvoord et al., 2008</td>
<td>Burke et al., 2006; Coppola and Daniels, 1996; Poock et al., 2007</td>
<td>Bommaraju, 2004; Scoles and Millan, 2005</td>
<td>Brod et al., 2010</td>
<td></td>
</tr>
<tr>
<td><strong>Advanced</strong></td>
<td>Nekvasil, 1998; Poronnik and Moni, 2006</td>
<td>Lillig, 2008; May et al., 2010; Reilly and Strickland, 2010</td>
<td>Alaimo et al., 2009; Lillig, 2008; May et al., 2010; Stoller et al., 2005</td>
<td>Advaned: Barr, 1995</td>
<td></td>
</tr>
<tr>
<td><strong>Capstone</strong></td>
<td></td>
<td></td>
<td>Capstone: Berry and Carlson, 2010</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scientific method</th>
<th>Biology/Life Sciences</th>
<th>Chemistry</th>
<th>Engineering</th>
<th>Math/Computer Science/Statistics</th>
<th>Physics/Earth Sciences</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Introductory</strong></td>
<td>Gerdeman et al., 2007</td>
<td>McClure, 2009</td>
<td>Bommaraju, 2004</td>
<td>Brod et al., 2010</td>
<td></td>
</tr>
<tr>
<td><strong>Advanced</strong></td>
<td>Clase et al., 2010</td>
<td>Alaimo et al., 2009; Lillig, 2008; May et al., 2010; Stoller et al., 2005</td>
<td></td>
<td>Capstone: Blakeslee, 1997</td>
<td></td>
</tr>
<tr>
<td><strong>Critical thinking</strong></td>
<td>MacKay et al., 2005; Pelaez, 2002</td>
<td>Burke et al., 2006; Coppola and Daniels, 1996; Poock et al., 2007</td>
<td>Alaimo et al., 2009; Kim et al., 2005; Lillig, 2008; May et al., 2010; Reilly and Strickland, 2010; Stoller et al., 2005</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Introductory</strong></td>
<td>Clase et al., 2010; Nekvasil, 1998; Ranelli and Nelson, 1998; Ryan and Campa, 2000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Advanced</strong></td>
<td>Reynolds and Thompson, 2011</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Continued)
### Table 2. Continued

<table>
<thead>
<tr>
<th>Biology/Life Sciences</th>
<th>Chemistry</th>
<th>Engineering</th>
<th>Math/Computer Science/Statistics</th>
<th>Physics/Earth Sciences</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Effective communication</strong></td>
<td><strong>Introductory:</strong> Coppola and Daniels, 1996; Forbes and Davis, 2008; Kovac and Sherwood, 2001; McClure, 2009; Robinson, 2008; Tilstra, 2001</td>
<td><strong>Introductory:</strong> Hanson and Williams, 2008; Newcomer et al., 2003; Yoder and Sawyers, 2006</td>
<td><strong>Introductory:</strong> Fleron and Hotchkiss, 2001</td>
<td><strong>Advanced:</strong> Blakeslee, 1997</td>
</tr>
<tr>
<td><strong>Advanced:</strong> Clase et al., 2010; Poronnik and Moni, 2006; Ranelli and Nelson, 1998; Ryan and Campa, 2000</td>
<td><strong>Advanced:</strong> Alaimo et al., 2009; Kim et al., 2005; Lillig, 2008; May et al., 2010; Robinson and Stoller, 2008; Sivey and Lee, 2008; Stoller et al., 2005</td>
<td><strong>Advanced:</strong> Craig et al., 2008; Hecker, 1997; House et al., 2007; Kim et al., 2005; Newcomer et al., 2003; Troy et al., 2004</td>
<td><strong>Introductory:</strong> Clase et al., 2010</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Capstone:</strong> Schepmann and Hughes, 2006; Wllmer and Latosi-Sawin, 1999</td>
<td><strong>Capstone:</strong> Mirel and Olsen, 1998; Ostheimer and White, 2005; Wojahn et al., 2001</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Metacognition</strong></td>
<td><strong>Introductory:</strong> Armstrong et al., 2008; MacKay et al., 2005</td>
<td><strong>Introductory:</strong> Hanson and Williams, 2008</td>
<td><strong>Introductory:</strong> Lerch et al., 2006</td>
<td><strong>Introductory:</strong> Goldberg and Bendall, 1995; Jang, 2007; Klein, 2004</td>
</tr>
<tr>
<td><strong>Introductory:</strong> Clase et al., 2010</td>
<td><strong>Advanced:</strong> Jang, 2007</td>
<td><strong>Advanced:</strong> Dahm et al., 2006; Thompson et al., 2005</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>Introductory:</strong> Mirel and Olsen, 1998</td>
<td><strong>Introductory:</strong> Fleron and Hotchkiss, 2001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Advanced:</strong> Clase et al., 2010</td>
</tr>
<tr>
<td><strong>Professionalization</strong></td>
<td><strong>Introductory:</strong> Gerdeman et al., 2007</td>
<td><strong>Introductory:</strong> Newcomer et al., 2003</td>
<td><strong>Introductory:</strong> Fleron and Hotchkiss, 2001</td>
<td><strong>Advanced:</strong> Blakeslee, 1997; Kelly et al., 2000</td>
</tr>
<tr>
<td><strong>Introductory:</strong> Reynolds and Thompson, 2011</td>
<td><strong>Advanced:</strong> Coppola and Daniels, 1996; Forbes and Davis, 2008; Robinson, 2008</td>
<td><strong>Advanced:</strong> Craig et al., 2008; Dahm et al., 2009; Dannels, 2000; Hecker, 1997; Newcomer et al., 2003; Troy et al., 2004</td>
<td><strong>Capstone:</strong> Mirel and Olsen, 1998</td>
<td><strong>Capstone:</strong> Fleron and Hotchkiss, 2001</td>
</tr>
<tr>
<td></td>
<td><strong>Advanced:</strong> Alaimo et al., 2009; Lillig, 2008; May et al., 2010; Reilly and Strickland, 2010; Robinson, 2008; Robinson and Stoller, 2008; Stoller et al., 2005</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Capstone:</strong> Wallner and Latosi-Sawin, 1999</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Gaps in the table do not necessarily indicate an absence of studies, but rather indicate that we did not identify illustrative studies in those areas. The database of 324 journal articles, books, book sections, conference proceedings, and reports (located at http://bit.ly/fjudgo) can be searched by key words (Table 3). To search for multiple key words simultaneously, use the advanced search feature of the database and specify search fields as “anywhere.”*
and therefore serve as potential intervention targets for WTL practices. Cognition involves the skills to encode and recall information: rehearsal, elaboration, organization, and comprehension-monitoring learning strategies (Weinstein and Mayer, 1985); and the processes of problem solving and critical thinking (Schraw et al., 2006). Metacognition involves planning, monitoring, and evaluating one’s cognitive processes. Motivation involves those prior beliefs and attitudes that affect engagement with the task and the development and use of cognitive and metacognitive processes (Schraw et al., 2006). Emotion involves anxiety associated with performance, for example test anxiety or stereotype threat, and the notion of “troublesome knowledge,” that is, when learning involves transformations in beliefs, commitments, and matters of identity (Meyer and Land, 2005). Although several studies have looked at the impact of metacognition in writing to promote learning gains (Thompson et al., 2005; Armstrong et al., 2008, Hanson and Williams, 2008), mechanisms of effect are rarely considered in WTL research in STEM (although Shah et al., 2009, is a notable example of how mechanisms could be studied).

Fourth, the extant evidence supports the effectiveness of two types of WTL assignments in particular for improving learning in STEM disciplines: 1) Assignments that focus critical reflection on one’s epistemic beliefs regarding knowledge and understanding, problem solving, and application of knowledge (e.g., Bangert-Drowns et al., 2004, Lerch et al., 2006); and 2) assignments that engage the student in formulating a reasoned argument (e.g., Kelly et al., 2000; Bradley, 2001; Kelly and Takao, 2002; Lerner, 2007; Armstrong et al., 2008).

Fifth, we urge the adoption of a “hybrid” research paradigm that builds on the insights, methods and rubrics, and interpretative frameworks that characterize WTL scholarship in the humanities and social sciences, while promoting the hypothesis testing, controls, and experimental paradigm typical of the cognitive and natural sciences (Van Maanen, 1988; Lave and Wenger, 1991; Kirsch and Sullivan, 1992; Hendlisman et al., 2004; Schell and Rawson, 2010). Such a “hybrid paradigm” would encourage multifactorial analytical and experimental-level studies that investigate and compare the impact of WTL practices on disciplinary-specific learning outcomes, using qualitative as well as quantitative assessment methods, as a function of hypothesized mediating and moderating variables, including emotional and motivational factors and learning context.

Finally, to address the gap between research and practice, we recommend that, in reporting on their work, researchers give attention to the kind of classroom situations and goals for which a specific WTL strategy is intended. Such information is necessary if practitioners are going to be confident in their choice and implementation of an intervention.

Rivard’s conviction that “The area of writing to learn in science is ideal for developing collaborative projects in classroom inquiry” (p. 976) remains as true today as it was 18 yr ago. What is different is that the combination of the emerging community of WTL/STEM educators, a learning-based conceptual framework, the database resulting from our heuristic review, and the adoption of the hybrid paradigm enables and empowers collaborative multi-university initiatives involving multidisciplinary teams of investigators to formulate and implement common protocols across multiple settings.

ACKNOWLEDGMENTS

We thank all the participants of our WTL workshop, and especially the other members of our working group: Greg Bothun, David Hanson, Jeffery Kovac, Lisa McNair, Tamara Moore, Marie Parette, Arlene Russell, and Leslie Schiff. We also thank Leo Gafney for acting as the external evaluator for the WTL workshop, Annaliese Franz for suggesting valuable references for the database, and Teddy Gray for

Table 3. Key words used to organize and search WTL in STEM bibliography database

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Research level</th>
<th>Assignment type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biology and Life Sciences</td>
<td>Descriptive</td>
<td>CPR</td>
</tr>
<tr>
<td>Chemistry</td>
<td>Analytic</td>
<td>Grant proposals</td>
</tr>
<tr>
<td>Engineering</td>
<td>Experimental</td>
<td>Group project</td>
</tr>
<tr>
<td>Math, Computer Science, and Statistics</td>
<td>Meta-analysis</td>
<td>In-class writing</td>
</tr>
<tr>
<td>Physics and Earth Science</td>
<td>Review</td>
<td>Journal articles (includes journal-style papers)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level</th>
<th>Research methods</th>
<th>Research methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introductory</td>
<td>Assessment</td>
<td>Assessment</td>
</tr>
<tr>
<td>Advanced</td>
<td>Analysis of assignments</td>
<td>Comparison groups</td>
</tr>
<tr>
<td>Capstone</td>
<td>Discourse analysis</td>
<td>Discourse analysis</td>
</tr>
<tr>
<td></td>
<td>Evaluations</td>
<td>Evaluations</td>
</tr>
<tr>
<td></td>
<td>Exams</td>
<td>Exams</td>
</tr>
<tr>
<td></td>
<td>Focus groups</td>
<td>Focus groups</td>
</tr>
<tr>
<td></td>
<td>Grounded analysis</td>
<td>Grounded analysis</td>
</tr>
<tr>
<td></td>
<td>Interviews</td>
<td>Interviews</td>
</tr>
<tr>
<td></td>
<td>Qualitative</td>
<td>Qualitative</td>
</tr>
<tr>
<td></td>
<td>Quantitative analysis</td>
<td>Quantitative analysis</td>
</tr>
<tr>
<td></td>
<td>Rubric</td>
<td>Rubric</td>
</tr>
<tr>
<td></td>
<td>Survey</td>
<td>Survey</td>
</tr>
<tr>
<td>Learning outcomes</td>
<td>Think-aloud protocols</td>
<td>Think-aloud protocols</td>
</tr>
<tr>
<td>Conceptual understanding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scientific method</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Critical thinking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Communication</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metacognition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Professionalization (includes “disciplinary ways of knowing”)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
help in synchronizing our database with a stable, searchable online database. This work was funded in part by NSF grant 000215159.

REFERENCES


Coppola BP, Daniels DS (1996). The role of written and verbal expression in improving communication skills in an undergraduate chemistry program. Language Learning across Disciplines 1, 67–86.


This study explores biology undergraduates’ misconceptions about genetic drift. We use qualitative and quantitative methods to describe students’ definitions, identify common misconceptions, and examine differences before and after instruction on genetic drift. We identify and describe five overarching categories that include 16 distinct misconceptions about genetic drift. The accuracy of students’ conceptions ranges considerably, from responses indicating only superficial, if any, knowledge of any aspect of evolution to responses indicating knowledge of genetic drift but confusion about the nuances of genetic drift. After instruction, a significantly greater number of responses indicate some knowledge of genetic drift \( (p = 0.005) \), but 74.6% of responses still contain at least one misconception. We conclude by presenting a framework that organizes how students’ conceptions of genetic drift change with instruction. We also articulate three hypotheses regarding undergraduates’ conceptions of evolution in general and genetic drift in particular. We propose that: 1) students begin with undeveloped conceptions of evolution that do not recognize different mechanisms of change; 2) students develop more complex, but still inaccurate, conceptual frameworks that reflect experience with vocabulary but still lack deep understanding; and 3) some new misconceptions about genetic drift emerge as students comprehend more about evolution.

INTRODUCTION

Biology educators have articulated the importance of teaching undergraduates the mechanisms of evolution. In a national survey, more than 300 college biology faculty agreed on the importance of evolution instruction in introductory biology sequences, including instruction on evolutionary mechanisms and phylogenetics (Gregory et al., 2011). In fact, evolution was the most agreed upon topic, with 89% of faculty agreeing it was an essential topic for biology students to learn. Similarly, in the report BIO2010, the Committee on Undergraduate Biology Education noted that students should understand that “all living things have evolved from a common ancestor through processes that include natural selection and genetic drift acting on heritable genetic variation” (National Research Council, 2003). Evolution is a core concept in
Misconceptions about Genetic Drift

In the present study, we used a mixed-methods approach to: 1) describe undergraduates’ definitions of genetic drift, 2) identify the most common misconceptions in those definitions, 3) examine differences in students’ definitions before and after receiving instruction on genetic drift, and 4) propose a framework for future research that interprets students’ misconceptions and illustrates how undergraduates’ understanding of genetic drift progresses.

METHODS

Our methodology was mixed. Our qualitative analytical methods aligned with grounded theory and were supplemented with statistical analysis to compare student responses before and after instruction.

In grounded theory, the central question is: “What theory emerges from systematic comparative analysis and is ground in fieldwork so as to explain what has been and is observed?” (Patton, 2002, p. 133). In practice, grounded theory aims to derive descriptions from the data, as opposed to approaching the data with preliminary explanations. Those data are read and re-read, and from these readings investigators establish categories that explain the data. Once categories are established, investigators review data again and assign units of data, such as quotes from student responses, to categories. Thus, categories serve to organize detailed descriptions of the data. In this way, grounded theory is analogous to inductive science, in which careful and repeated observations enable descriptions. Additionally, grounded theory, like inductive science, may produce hypotheses that can be tested with additional research (i.e., deductive science). Grounded theory was developed by sociologists and is traditionally used to analyze interview data with the goal of developing theories about human actions, interactions, and social processes (Creswell, 2007). In our study, we analyze qualitative data from written responses, focusing on participants’ conceptions, rather than the broader context in which the participants are acting.

For this study, we synthesized data collected during two distinct research projects (Table 1). Authors from a National Evolutionary Synthesis Center (NESCent) working group (T.M.A., R.M.P., L.S.M., T.L.M., A.T., and K.E.P.) collected data in preparation for the development of a concept inventory on genetic drift. Authors affiliated with the National Center for Case Study Teaching in Science (C.F.H., D.R.T., and P.P.L.) collected data during a study of the effectiveness of a series of case studies, including one on genetic drift and other evolutionary mechanisms. Combining data sets allowed us to analyze misconceptions about genetic drift from a broad range of students and to capitalize on the different strengths of each project (Table 1). The case study data set allowed us to test for differences between responses collected before and after instruction, whereas the concept inventory data set provided information about misconceptions that occur after more than one exposure to genetic drift instruction, since many of the participants were biology majors enrolled in courses for which introductory biology was a prerequisite. We describe methods used to collect both data sets in the Supplemental Material.
Expert ideas, we need to define how we used the term misconception in this study. We considered the term misconception to distinguish between ideas generated during data collection and deeply held ideas. We defined a misconception as a scientifically inaccurate idea about a scientific concept. These inaccuracies may occur before and after instruction. We did not distinguish between ideas generated during data collection and deeply held ideas. We considered the term misconception to be equivalent to the term alternative conception, and to be a particular kind of preconception or naïve conception (Gilbert and Watts, 1983).

In the literature on natural selection misconceptions, the term misconception is often defined more narrowly than we have defined it in this study. For example, natural selection misconceptions have been referred to as “deeply rooted” and as “intuitive interpretations of the world” (Cunningham and Wescott, 2009; Gregory, 2009). However, natural selection misconceptions have been explored in depth, leading to more precise definitions of natural selection misconceptions. In contrast, few, if any, studies have focused on students’ conceptions of genetic drift. We have used a broad definition of misconceptions that encompasses all students’ inaccurate ideas, because considerably more research will be necessary to identify which inaccurate ideas are intuitive, common across diverse populations, and deeply held.

Rigor in qualitative research has been defined as the “attempt to make data and explanatory schemes as public and replicable as possible” (Norman Denzin, as quoted in Anfara et al., 2002, p. 7). Therefore, two authors followed this systematic approach:

1. We (T.M.A. and P.P.L.) independently identified student misconceptions about genetic drift in the concept inventory data set and case study data set, respectively. Thereafter, we combined data sets and completed all analyses in the same place and time, which allowed us to immediately deliberate on any ambiguities.
2. We agreed on an initial list of misconceptions about genetic drift. To create this initial list, we analyzed a subsample of student responses from the combined data set to establish the characteristics of misconceptions.
3. We each coded ~40 responses, identifying the misconceptions in each response to establish that we could reliably classify misconceptions. We discussed any discrepancies until we reached consensus. At this preliminary stage, we identified three general types of student responses: responses that did not address genetic drift, despite explicit instructions to do so; responses containing misconceptions about genetic drift; and responses indicating at least some knowledge of genetic drift. These general types were not mutually exclusive.
4. Using the initial list of misconceptions produced in step 3, we began coding the full data set. Any idea we could not classify after discussion was coded as undetermined. Coding the full data set was necessarily iterative. Throughout this process, new misconceptions emerged, our descriptions of existing misconceptions were refined and sometimes subdivided, and the data were recoded accordingly. In all cases, new misconceptions were closely related to misconceptions from our initial list, so it was only necessary to reanalyze responses previously coded as containing a misconception closely related to the newly emerged misconception and responses previously coded as undetermined.
5. After all responses had been analyzed and coded at least once, we re-read all of the responses containing the same misconception, and discussed at length the characteristics delineating each misconception.
6. Toward the end of our analysis, we tested our list of misconceptions to ensure it was exhaustive. We drew a new, random sample of 30 responses from the case study project data set, including questionnaires completed before and after instruction from all six sections, and coded this sample. We found no misconceptions we could not classify with our final coding system. We therefore concluded our list of misconceptions included all but the rarest genetic drift misconceptions held by participating students.

As we analyzed the data, we looked for overarching categories that would enable us to build a framework for future research on students’ conceptions about genetic drift. This is the end product of a grounded theory study (Creswell, 2007; Glaser and Strauss, 2010). We designed the framework to facilitate the interpretation of undergraduates’ misconceptions about genetic drift and to hypothesize how undergraduates’ understanding of drift may progress. To build the framework, three investigators (T.M.A., R.M.P., and P.P.L.) iteratively grouped the full set of misconceptions and named the resulting clusters. We worked to propose a final framework that was derived from the data, not from explanations about student conceptions that we held prior to data analysis. This process continued until all three investigators agreed that the framework was true to the data and suggested testable hypotheses about genetic drift.

Statistical Analyses

We used 319 responses collected as part of the case study project to examine differences in students’ conceptions about genetic drift before and after instruction. As described in the Supplemental Material, we used a systematic sampling design to select student responses. We sampled different students’ responses before and after instruction, even though this precluded using matched pairs. This approach, which allowed us to include more students, ensured a breadth of responses, even though it limited statistical power by...
not controlling for individual variation among students. All statistical analyses were conducted in R, an open-source statistical analysis program (R Project for Statistical Computing, 2011).

To assess the hypothesis that instruction improved students’ understanding of genetic drift, we tested three predictions generated by this hypothesis. We predicted that before and after instruction the following would be different:

1. Number of students who did not address drift
2. Number of responses that indicated some knowledge of the definition of genetic drift
3. Number of responses containing at least one misconception

We tested these predictions with Fisher’s exact tests (Ramsey and Schafer, 2002); for predictions 2 and 3, we excluded responses that did not address drift. The Fisher’s exact test is more precise than a chi-squared test when some cells in a contingency table have small sample sizes (Ramsey and Schafer, 2002). A small p value resulting from this test indicates that the counts of responses in the two categories are not independent. In other words, a small p value resulting from the tests described above would suggest that instruction influenced students’ responses.

Finally, we used descriptive statistics to examine differences between the frequency of misconceptions before and after instruction in introductory biology courses and among upper-division students. We did not pursue additional statistical analysis for individual misconceptions or categories of misconceptions, as there were small sample sizes for some misconceptions and a lack of independence among groups resulting from responses containing more than one misconception.

RESULTS

Out of 356 student responses analyzed, few defined or attempted to apply the concept of genetic drift without using misconceptions. Even though questions from both data sets specifically asked students to define genetic drift, 31.5% (n = 112) of responses failed to address drift at all (Table 2). Among responses that addressed drift (n = 244), only 11.5% (n = 28) indicated some knowledge of the definition of genetic drift. Overall, 83.2% (n = 203) of the responses that addressed drift contained at least one misconception (Table 2). Some responses (n = 25) hinted at knowledge of genetic drift (e.g., included the term random or chance), but were too vague to be fully evaluated. Note that, because some responses indicated knowledge of genetic drift but also contained misconceptions, the percentages provided here sum to greater than 100%.

Categories of Student Misconceptions Regarding Genetic Drift

In responses that addressed drift (n = 244), we identified five overarching categories of misconceptions: Novice Genetics, Novice Evolution, Associating Genetic Drift with Other Evolutionary Mechanisms, Associating Genetic Drift with Population Boundaries, and Developing Genetic Drift Comprehension. These overarching categories are further divided into 16 distinct misconceptions that we describe below and summarize in Table 3. We also describe the frequency of each misconception (Table 3). We further divide the frequency of each misconception into those collected before and after introductory genetic drift instruction (case study data set) and those collected from students enrolled in upper-division biology courses (concept inventory data set) (Table 3).

Our detailed description of the misconceptions begins with the most novice overarching categories (Novice Genetics and Novice Evolution) and concludes with the most advanced category (Developing Genetic Drift Comprehension). The two categories presented in the middle (Associating Genetic Drift with Other Evolutionary Mechanisms, Associating Genetic Drift with Population Boundaries) do not represent a progression; rather, some responses in each category range from novice to developing comprehension. Within the overarching categories of misconceptions, we have listed misconceptions in decreasing order from highest to lowest percentage of responses that addressed drift (Table 3). It is important to recognize that although some misconceptions we describe indicated more advanced knowledge than others, responses in the most advanced category still differ in key ways from an expert’s conception of genetic drift.

We use quotes from students to illustrate the misconceptions encompassed by each overarching category. In the interest of brevity, we include the most salient sections of a response, rather than complete responses. In some cases, we may have used additional information included in a response.
<table>
<thead>
<tr>
<th>Misconceptions</th>
<th>Student quotes</th>
<th>% of Total (n = 244)</th>
<th>% Before instructionb (n = 85)</th>
<th>% After instructionb (n = 122)</th>
<th>% Upper divisionc (n = 37)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Novice Genetics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genetic drift is... shared traits or genes.</td>
<td>“Genetic drift [is] when it’s the same species but different characteristics.”</td>
<td>12.7</td>
<td>22.4</td>
<td>9.0</td>
<td>5.4</td>
</tr>
<tr>
<td></td>
<td>“Genetic drift because both species [have] distinctive commonalities.”</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>“Genetic drift is where the amount of present alleles change[s] gradually over time.”</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>“Genetic drift is a change in genes over time.”</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>“Genetic drift is the passing down of traits while natural selection does not have anything to do with genetics.”</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Novice Evolution</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genetic drift is... acclimation to the environment that may result from a need to survive.</td>
<td>“It was probably genetic drift. As the butterflies adapted to their new habitat they had to physically change in order for survival.”</td>
<td>20.9</td>
<td>31.8</td>
<td>14.7</td>
<td>13.5</td>
</tr>
<tr>
<td></td>
<td>“The evolution of the two butterflies is genetic drift because they developed to their surroundings.”</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>“[Genetic drift occurred when] certain butterflies with each gene and characteristics came together in a certain spot and they mated forming new types of butterflies.”</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>“[Genetic drift is] the genetic changes that occur when a population is not under selection.”</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Associating Genetic Drift with Other Evolutionary Mechanisms</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genetic drift is... random mutation.</td>
<td>“[Genetic drift occurs when] due to random mutations, genetic structure can change over time.”</td>
<td>18.8</td>
<td>13.0</td>
<td>13.1</td>
<td>48.6</td>
</tr>
<tr>
<td></td>
<td>“The definition of genetic drift is random chance mutation.”</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>“The movement of genes from one population of a species to another or from one locality to another.”</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>“Genetic drift is a chance occurrence that brings genes into a population.”</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>“Genetic drift occurs to eliminate the less adaptable trait that is not well suitable to the environment.”</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>“[Genetic drift is] the process of changing allele frequencies within a population.”</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Associating Genetic Drift with Population Boundaries</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genetic drift is... migration with or without acclimation to the environment.</td>
<td>“Genetic drift is when the population moves to a location more suitable to its characteristics.”</td>
<td>32.8</td>
<td>33.0</td>
<td>36.1</td>
<td>21.6</td>
</tr>
<tr>
<td></td>
<td>“[Genetic drift occurred] as certain ancestral butterflies moved to different areas, they changed to better suit their new environment.”</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Continued
Table 3. Continued

<table>
<thead>
<tr>
<th>Misconceptions</th>
<th>Student quotes</th>
<th>% of Total (n = 244)</th>
<th>% Before instruction&lt;sup&gt;b&lt;/sup&gt; (n = 85)</th>
<th>% After instruction&lt;sup&gt;b&lt;/sup&gt; (n = 122)</th>
<th>% Upper division&lt;sup&gt;c&lt;/sup&gt; (n = 37)</th>
</tr>
</thead>
</table>
| the separation of populations with or without acclimation to the environment. | “[Genetic drift occurs due to] isolation of a population or species by whatever means.”  
“Genetic drift occurs when a sect of a species is separated from the other and changes to adapt to their new environment.”  
“I believe [it was genetic drift] because I believe at one point both species were one, then separated.”  
“It was genetic drift because some genes changed to create this new species.” | 10.2 9.4 10.7 10.8 | 9.4 9.8 10.7 10.8 |                                                                 |                                                                                   |
| speciation.                                                                    |                                                                                                                                                                                                               |                     |                                         |                                        |                                    |
| Developing Genetic Drift Comprehension                                          | Genetic drift is... a change in genes caused by an isolated event, often a catastrophe. limited to small populations, when an allele is fixed in a population.                                           | 8.6 0.0 12.3 18.9   | 0.0 0.0 8.2 2.7                       | 2.5 0.0 8.1 8.1                                                   | 1.6 0.0 0.8 8.1                                                                    |
|                                                                 | “Genetic drift involves a natural disaster that dramatically changes the genes in that area.”  
“Genetic drift is genetics in a smaller populations.”  
“This is when alleles from one population either die out or become the only allele present. It occurs because of random processes. The alleles just happen to die out or become the most prevalent because of chance.” |                     |                                         |                                        |                                    |

<sup>a</sup>Frequencies are based on the subset of responses that addressed drift (n = 244), not the total number of responses (n = 356).

<sup>b</sup>Responses from the case study project.

<sup>c</sup>Responses from the concept inventory project.

Some responses in Novice Genetics vaguely described genetic drift as gradual genetic change in a population without describing a mechanism of change:

“Genetic drift = gradual change in genes.”

“[This is genetic drift because] their similar characteristics indicate that over time the genetics of the species slowly changed.”

A few responses in Novice Genetics defined genetic drift as occurring when genes or traits are passed from one individual to another. Responses were not always specific about the units between which traits or genes were passed. Some described genes passing from parent to offspring through reproduction, but others described the transmission of traits between individuals:

“Genetic drift is when certain desirable characteristics that may occur through mutation are passed on to offspring.”

“Genetic drift is the flow of genes from one individual to another.”

**Category 2: Novice Evolution.** Responses in the Novice Evolution category defined genetic drift as an evolutionary mechanism but conflated the definition of genetic drift with novice conceptions of evolution. The answers indicated little or no knowledge of random occurrences. The most common misconception in Novice Evolution has also been identified and
described in studies of students’ misconceptions regarding natural selection (e.g., Bishop and Anderson, 1990; Nehm and Reilly, 2007). These responses defined genetic drift as the process, or result, of the environment causing change over time, attributing this change to "adaptation," by which they seemed to mean acclimation to environmental characteristics. Some responses containing this misconception explicitly stated that change resulted from a need to survive:

- "Genetic drift is the most reasonable answer because the sun brings out brightness like the bright butterfly and the shade is dark like the darker butterfly."
- "Genetic drift is when a species changes due to a specific need to survive or thrive."

Another misconception in Novice Evolution defined genetic drift as an evolutionary mechanism in which change results from mating between individuals from different species:

- "The butterflies were the same color and liked the same environments but began breeding with butterflies of different kinds, possibly because of food scarcity or wind currents."
- "Genetic drift is change due to breeding."

Lastly, a few responses in Novice Evolution contained the misconception that genetic drift is a mechanism of evolutionary change that occurs when natural selection cannot or is not occurring. The descriptions in these responses were so superficial that despite the use of key terms like natural selection, the responses failed to indicate any understanding of evolutionary processes. This misconception was not common, but was very clearly articulated in two responses collected from students in different courses in response to different questions:

- "[Genetic drift is] the genetic changes that occur when a population is not under selection."

Category 3: Associating Genetic Drift with Other Evolutionary Mechanisms. Biologists recognize natural and sexual selection, mutation, gene flow, and genetic drift as distinct evolutionary mechanisms. Responses in Associating Genetic Drift with Other Evolutionary Mechanisms confused genetic drift with other evolutionary mechanisms or with evolution in general. The definitions in these responses indicated developing comprehension of evolution, but did not indicate knowledge of genetic drift.

The most common misconception in Associating Genetic Drift with Other Evolutionary Mechanisms defined genetic drift as random mutation. About half of these responses explained that genetic drift results from mutations, while the other half defined genetic drift as the process of mutation or the accumulation of mutations over time. In some cases, students specified a precise mechanism of mutation:

- "Genetic drift = change in a population due to mutation."
- "Genetic drift is the drifting of genes during mutations. A base pair is usually cutoff, that alters the gene sequence leading to changed genes."

Another misconception in this category defined genetic drift as gene flow. Specifically, these responses described genetic drift as the process of alleles entering or leaving populations or as the process of alleles from different populations "mixing." Some responses described the movement of genes, rather than the movement of alleles. Notably, Nehm and Reilly (2007) identified this misconception in undergraduates’ responses to an open-response item designed to measure knowledge of natural selection:

- "Genetic drift involves the movement of alleles out of populations/gene pools to new environments."
- "Gene exchange between different populations of animals. Results in an increase or decrease of a specific type of gene."

The third misconception in Associating Genetic Drift with Other Evolutionary Mechanisms defined genetic drift as natural selection. In some cases, these definitions of natural selection were nuanced and accurate; in other cases, responses were less detailed, but implied or described an interaction between traits and the environment resulting in differential reproductive success, survival, or fitness. One response defined genetic drift as sexual selection:

- "Genetic drift occurs because survival of the fittest so if some alleles that are passed down to offspring provide a benefit, those alleles are more likely to get passed on to their offspring."
- "Genetic drift is the gradual change in the frequency of specific alleles in a population to be more or less common [and]...occurs when there is a change in the environment that makes specific traits more or less favorable for fitness."

Finally, a few responses in this category defined genetic drift as any change in allele frequencies:

- "Genetic drift is when there is a change in the allele frequency of a population. ‘Drift is the alteration of genes by anything, including chance.’"

Category 4: Associating Genetic Drift with Boundaries between Populations. Biologists recognize the founder effect to be one scenario in which genetic drift can occur. Essentially, when a small random sample of individuals from a larger population become the founders of a new population, they are likely to carry only a fraction of the genetic variation of the original population (Futuyma, 2005). Additionally, founding populations are often small and are therefore likely to be further impacted by genetic drift for many generations following the founding event. Moreover, genetic drift and natural selection can lead to reproductive isolation in a peripheral population, such as a founding population. This process is called peripatric speciation (Futuyma, 2005). No responses in Associating genetic drift with boundaries between populations came close to indicating knowledge of the nuanced concepts just described. However, these responses defined genetic drift as movement, separation, and/or speciation, which hinted at knowledge of, or at least exposure to, founder effect as an example of genetic drift.
The most common misconception in Associating Genetic Drift with Population Boundaries defined genetic drift as migration, by which responses typically seemed to mean emigration. In some cases, responses described migration followed by adaptation to the environment. The descriptions of adaptation in these responses were similar to those in Novice Evolution in that they described adaptation as acclimation to environmental characteristics, but these responses were distinct in that they also discussed the movement of individuals. The units discussed in these responses included individuals, species, and populations. Some responses discussed individuals or groups moving to locations better suited to their traits.

The terms migration and gene flow are often used interchangeably by experts, who recognize that migration is an evolutionary process only when it leads to a change in allele frequencies. There was no indication that students understood this subtlety. The responses in this category differed from those that defined genetic drift as gene flow, because they did not mention the movement of alleles or genes:

“Genetic drift is when a certain species migrates to another location.”

“Genetic drift would be where members of a population with different traits move to an environment that fits these traits.”

“Genetic drift would take place if the butterflies would have migrated to another climate and adapted to their surroundings by the means of migrations.”

A similar misconception defined genetic drift as the separation or isolation of populations. In some cases, responses discussed separation followed by adaptation, by which they seemed to mean acclimation, to a new environment:

“Genetic drift generally happens when part of a species population is separated and becomes distinguished and change[s].”

“Genetic drift is when members of the same species get separated by environmental forces and over time develop differently.”

The third misconception in this category defined genetic drift as speciation. While it is possible for genetic drift to contribute to speciation, these responses did not provide an explanation for how speciation would occur. About half of these responses defined genetic drift as speciation following the separation of populations:

“These species of butterflies were once the same then slowly over time began shifting into one species that prefer sunny meadows and another that prefers dense woodlands.”

“Genetic drift occurs when an offshoot of a population starts to develop traits that separate it from the original population, usually by a chance act.”

“Genetic drift happens when two species become isolated from each other or no longer reproduce, creating a cross breeds.”

Category 5: Developing Genetic Drift Comprehension. Biologists recognize many nuances of the process of genetic drift.

For example, genetic drift can result from random sampling of gametes during sexual reproduction, as well as random sampling of individuals, and their gametes, resulting from a population bottleneck (Futuyma, 2005). Experts recognize drift occurs in all finite populations, but is likely to have a more pronounced impact given a small effective population size (Barton et al., 2007). Experts also know genetic drift can, but does not always, lead to the fixation of alleles, and that genetic drift tends to decrease genetic variation within a population and increase variation among populations (Frankham et al., 2002).

Responses in Developing Genetic Drift Comprehension indicated some knowledge of genetic drift. However, the definitions in this category placed inaccurate limitations on the circumstances under which genetic drift can occur.

The most common misconception in Developing Genetic Drift Comprehension defined genetic drift as, or as resulting from, an isolated event, often a catastrophe. These responses did not recognize genetic drift as a process occurring each generation:

“Genetic drift is where there is some event that decreases the variation in a population.”

Another misconception in Developing Genetic Drift Comprehension limited genetic drift to small populations:

“Genetic drift is a change in allele frequency due to a random genetic occurrence in a small population.”

The least common response in Developing Genetic Drift Comprehension described genetic drift as allele fixation, rather than describing fixation as a potential result of genetic drift:

“[Genetic drift is] when an allele gets fixed on a population.”

“[Genetic drift is] allele fixation due to limited gene pool.”

“[Genetic drift is when] a random event knocks out one genotype.”

Vague Responses That Hinted at Knowledge of Genetic Drift

Responses that hinted at knowledge of genetic drift used terms such as random or chance but otherwise did not indicate knowledge of genetic drift. In some cases, the term random or chance was embedded in misconceptions, but in most cases these responses were simply too vague to evaluate:

“Genetic drift is all about chances to the outcome of the offspring.”

Responses Indicating Some Knowledge of Genetic Drift

Responses indicating some knowledge of drift ranged considerably in quality. Some responses provided precise and nuanced definitions of genetic drift, others gave brief but accurate descriptions of drift, and some responses included misconceptions.

The following quote was one of the most articulate responses in our sample. In particular, the subtle and precise
language differentiates this response from responses containing misconceptions. Though the response discusses an event or catastrophe leading to genetic drift—like responses in Developing Genetic Drift Comprehension—the use of the introductory clause “for instance” suggests that the student recognizes this is one example of drift, rather than the only circumstance under which drift takes place:

“Genetic drift is evolution that occurs purely by chance. For instance, an F1 generation could have 10 red flowers, 10 pink flowers, and 10 white flowers. If all the white flowers are accidentally killed or something happens, their genes will not be passed on to future generations.”

In contrast, the next quote demonstrates how a response can indicate some knowledge of genetic drift and contain a misconception. The first sentence of the response confuses genetic drift with selection, while the second sentence indicates knowledge of genetic drift:

“Genetic drift occurs when through sexual or natural selection, certain alleles are favored. Additionally, it may just so happen that an allele becomes more or less prevalent though it neither helps nor harms individuals within a population.”

Results of Statistical Analyses

We used statistical analyses to address three predictions about student learning. We tested these predictions using data from the case study project (n = 319), because this project collected data before and after introductory-level genetic drift instruction. We predicted that 1) the number of students who did not address drift, 2) the number of responses that indicated some knowledge of the definition of genetic drift, and 3) the number of responses containing at least one misconception would all be different before and after instruction.

All three of the predictions about student learning were supported by our data (Table 2). In all cases, students exhibited more knowledge of genetic drift after instruction. The number of responses that did not address drift was significantly different before and after instruction (Fisher’s test, p < 0.0001; Table 2), suggesting that students in these courses did not address drift before instruction because they had little or no knowledge of the concept. To test our second and third predictions, we examined only the responses from the case study data set in which students addressed drift (n = 207). The number of responses indicating some knowledge of genetic drift was different before and after instruction (p = 0.005; Table 2). Additionally, the number of responses containing at least one misconception was different before and after instruction (p < 0.0001; Table 2).

When we examined the frequency of student responses containing each of the 16 distinct misconceptions at different stages of instruction, we noticed that while some misconceptions were less common among students who had received genetic drift instruction, other misconceptions were more common following instruction (Table 3). Specifically, the misconceptions in Novice Genetics and Novice Evolution were less common after introductory instruction and among upper-division students, whereas misconceptions in Developing Genetic Drift Comprehension were absent before instruction, but increasingly common with more instruction (Table 3). The frequency of misconceptions in Associating Genetic Drift with Other Evolutionary Mechanisms and Associating Genetic Drift with Population Boundaries remained about the same before and after introductory instruction, but among upper-division students these two categories diverged (Table 3). Misconceptions in Associating Genetic Drift with Other Evolutionary Mechanisms were substantially more common among upper-division students than among introductory students, whereas misconceptions in Associating Genetic Drift with Population Boundaries were less common among upper-division students than among introductory students (Table 3).

DISCUSSION

Our observations represent the first effort, to our knowledge, to describe students’ conceptions of genetic drift and how those conceptions change over time. Among students who addressed genetic drift in their responses, nearly all (99%) undergraduates in introductory biology courses had misconceptions about genetic drift before instruction, and almost 75% retained misconceptions after explicit genetic drift instruction (Table 2). Furthermore, undergraduates who had completed introductory biology and were enrolled in upper-division biology courses for biology majors still had serious misconceptions about genetic drift (Table 3).

To facilitate future research on student conceptions of genetic drift, we propose a framework to interpret students’ conceptions about genetic drift and to describe how those conceptions change as students learn. This framework suggests three hypotheses regarding undergraduates’ conceptions of genetic drift. The rest of this paper presents the framework and hypotheses, followed by implications for instruction and future research.

Framework

Our framework includes the five broad categories of misconceptions identified during our qualitative analysis. The arrows between categories of misconceptions in our framework represent ways in which students’ conceptions may be changing as they learn (Figure 1). At one end of the framework are two categories of misconceptions most common among students before genetic drift instruction (Novice Genetics and Novice Evolution). Responses including misconceptions in these categories indicated no knowledge of genetic drift and only superficial—if any—knowledge of evolution. In the middle of the framework are two categories of misconceptions (Associating Genetic Drift with Population Boundaries and Associating Genetic Drift with Other Evolutionary Mechanisms) that were more common in students’ responses after some genetic drift instruction. These responses tended to use appropriate terminology about evolution, but did so in a way that revealed misconceptions and was often imprecise and disorganized. At the other end of the framework is the category of misconceptions indicating some knowledge of genetic drift, but also some confusion (Developing Genetic Drift Comprehension). Misconceptions in this category were most common among upper-division students who presumably had the most exposure to genetic drift.
I. Undeveloped conceptions

II. Confused conceptions

III. New misconceptions emerge

Figure 1. This framework hypothesizes how students’ conceptions of genetic drift change over time. Each circle represents an overarching category of misconceptions. Arrows represent the ways in which students’ conceptions may be changing as they learn. (I) Students enter introductory biology with undeveloped conceptions of evolution that do not distinguish among mechanisms of evolutionary change. (II) Students’ conceptual frameworks of evolution grow more complex, but are still highly inaccurate. (III) Students reject some misconceptions but form new ones regarding inaccurate constraints on when drift occurs.

We did not include a stage representing Expertise in Genetic Drift in our framework, because we derived our framework solely from our data. The standard for expertise would be for students to comprehend genetic drift without misconceptions and to correctly apply their comprehension to novel problems dealing with drift. Students in our data set did not demonstrate this level of expertise. For example, we asked the participants in the concept inventory study to explain experimental results using their knowledge of genetic drift and none were able to do so.

On the basis of framework, we propose three hypotheses regarding undergraduates’ conceptions of genetic drift. First, we hypothesize that most students enter introductory biology courses with an undeveloped conception of evolution that does not distinguish among mechanisms of evolutionary change (Figure 1, I). Common misconceptions documented in studies of students’ conceptions of natural selection were actually common misconceptions about genetic drift as well (Bishop and Anderson, 1990; Nehm and Reilly, 2007; Gregory, 2009). For example, students defined genetic drift as acclimation to the environment. The fact that these common misconceptions are associated with drift, as well as natural selection, suggests they are actually misconceptions about evolution in general. It appears that students who know nothing about genetic drift are using the context of the question or cues in class to associate genetic drift with evolution. They are then defining genetic drift as they would define evolution or natural selection, perhaps because they think all evolution is natural selection (Jakobi, 2010). If misconceptions in Novice Evolution do in fact become less common after instruction, as our data suggest (Table 3), that would support the hypothesis that students begin with a simplistic conception of evolution that grows more complex as they learn.

Second, we hypothesize that students’ conceptual frameworks of evolution grow more complex as they learn, but the added complexity is not necessarily more accurate than their previous, less complex, conceptual frameworks, nor is it expertly organized (Figure 1, II). Students seem to be gaining knowledge of biology vocabulary and concepts, but still lack deep understanding of concepts and scientifically accurate connections among concepts. Their definitions of genetic drift mix misconceptions, imprecise terminology, and irrelevant information with some accurate information. Responses containing misconceptions in the two categories at the center of our framework illustrate this confusion (Figure 1).

Student conceptions probably do not skip from the novice to the developing comprehension end of our framework, but instead must move through the muddled intermediate stage (Figure 1). The challenge for us as instructors is to move students through this stage effectively and efficiently, especially in introductory courses. An exciting area of future research will be to test the efficacy of teaching modules geared to addressing this issue.

Third, we hypothesize that genetic drift instruction leads to the rejection of some misconceptions and the formation of new ones (e.g., Yip, 1998). We observed that after instruction, fewer students had misconceptions in Novice Genetics and Novice Evolution, but more students had misconceptions in Developing Genetic Drift Comprehension (Figure 1, III). Among upper-division students, 48.6% had misconceptions in Associating Genetic Drift with Other Evolutionary...
Mechanisms. This is a substantially larger percentage than we observed among introductory students before (13.0%) or after (13.1%) instruction, suggesting additional genetic drift instruction revealed or generated misconceptions in this category. This result is simultaneously encouraging and discouraging. It is encouraging, because it indicates students’ ideas are changing. But it is discouraging in that most students still had misconceptions as upper-division biology undergraduates. It remains to be seen, through additional research, what conditions contribute to the development of Expertise in Genetic Drift (i.e., understanding and application without misconceptions).

Implications for Instruction
Our observations suggest that genetic drift is a challenging topic for students to learn. We have not found any exercises to teach genetic drift that have been assessed for impact on student learning, but a number of scholars have proposed ideas for teaching genetic drift and improving an instructor’s degree of comfort with the concept (e.g., Staub, 2002; Young and Young, 2003; Masel, 2011 [includes a description of the classic experiment by Peter Buri], 2012).

Though it remains unclear what strategies might effectively facilitate student learning of genetic drift, our observations indicate one potential problem to avoid. Instruction that provides limited examples of genetic drift in action may inadvertently teach students that drift occurs only in such cases. For example, if instruction focuses on the founder effect, students may extrapolate that genetic drift only occurs when individuals move from one location to another or when a subset of a population is isolated from the larger population. Alternatively, students may assume genetic drift only occurs in small populations when scenarios used in class focus exclusively on drift within small populations.

Implications for Future Research
Evidence is accumulating that the student misconception that need is a rationale for change is common across biology concepts. Though biological explanations including the term “need” are not necessarily illegitimate (Zohar and Ginossar, 1998), teleological reasoning commonly results in misconceptions. The most common misconception we observed among students was defining genetic drift as acclimation to the environment and, in many cases, describing acclimation as resulting from a need to survive. The most common misconception about natural selection is also the idea that individuals or populations change because they need to (Gregory, 2009). This misconception extends beyond conceptions of evolution as well. When asked to explain pictures of biological phenomena, such as a plant growing toward the sun or a group of birds flying in a V formation, the most common idea provided by elementary and secondary school students was that organisms changed because they needed to (Southerland et al., 2001). Adults who had taken multiple college-level science courses also commonly explained natural phenomena as existing to fulfill a need (Kelemen and Rosset, 2009). If this single (albeit tenacious) misconception is affecting students’ ability to learn concepts throughout biology, instruction specifically designed to help students think critically about this sort of reasoning could have an impressive impact on student learning. Future research can explicitly focus on determining the pervasiveness of the idea that need is a rationale for change in biological systems and on effective strategies for changing this misconception to a scientifically accurate explanation.

Future research is also necessary to fill out and refine our framework of how students learn genetic drift. Interviews will be valuable to gain deeper insight about student conceptions and how they change with instruction. A broader student population would also be valuable. For example, studying a larger sample of advanced undergraduates will be necessary to understand how student conceptions of genetic drift progress, including how instruction reveals or creates new misconceptions. Furthermore, different questions are likely to elucidate additional misconceptions (Nehm and Ha, 2011).

Finally, future research can document how experts define genetic drift, as well as outlining the key concepts and skills needed to demonstrate expertise in genetic drift. Genetic drift is fundamental to evolution, yet often overlooked. For example, Teaching about Evolution and the Nature of Science (National Academy of Sciences Working Group on Teaching Evolution, 1998) outlines the major themes in evolution (Ch. 2) but never mentions genetic drift. In the more recent Vision and Change (American Association for the Advancement of Science, 2011), evolution is included in the list of core concepts that all undergraduates should understand but genetic drift is hardly mentioned. To correct this oversight, genetic drift experts and biology education experts need to collaborate to describe what a student who has a complete understanding of genetic drift should be able to do with that knowledge. It would also be useful for experts to think about the necessary scaffolds for learning genetic drift, as well as the recommended timing of scaffolding, for example, in high school biology, undergraduate introductory biology, and undergraduate advanced biology. We have begun to address this aim by uncovering misconceptions about genetic drift among biology undergraduates, and future research on student conceptions of drift has the potential to be just as fruitful.

ACKNOWLEDGMENTS
This material is based on work supported by the National Science Foundation (NSF) under grant number DUE-0920264 awarded to the National Center for Case Study Teaching in Science. Additional support was provided by the National Evolutionary Synthesis Center (NSF grant number EF-0905606). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. We thank the students who participated in this study and their instructors. We also thank Erin L. Dolan, Steven T. Kalinowski, our reviewers, and the other members of the EvoCI Toolkit Working Group. Susan C. Alberts wrote the original draft of the butterfly question used as part of the case study project. This is a publication of both the University of Georgia Science Education Research Group and the EvoCI Toolkit Working Group at the NESCent.

REFERENCES
Misconceptions about Genetic Drift


Misconceptions about Genetic Drift
The C.R.E.A.T.E. Approach to Primary Literature Shifts Undergraduates’ Self-Assessed Ability to Read and Analyze Journal Articles, Attitudes about Science, and Epistemological Beliefs

Sally G. Hoskins,* David Lopatto,† and Leslie M. Stevens‡

*Department of Biology and The Graduate Center, City College of the City University of New York, New York, NY 10031; †Department of Psychology, Grinnell College, Grinnell, IA 50112; ‡Section of Molecular Cell and Developmental Biology, University of Texas at Austin, Austin, TX 78712

Submitted March 18, 2011; Revised June 9, 2011; Accepted July 5, 2011

The C.R.E.A.T.E. (Consider, Read, Elucidate hypotheses, Analyze and interpret data, Think of the next Experiment) method uses intensive analysis of primary literature in the undergraduate classroom to demystify and humanize science. We have reported previously that the method improves students’ critical thinking and content integration abilities, while at the same time enhancing their self-reported understanding of “who does science, and why.” We report here the results of an assessment that addressed C.R.E.A.T.E. students’ attitudes about the nature of science, beliefs about learning, and confidence in their ability to read, analyze, and explain research articles. Using a Likert-style survey administered pre- and postcourse, we found significant changes in students’ confidence in their ability to read and analyze primary literature, self-assessed understanding of the nature of science, and epistemological beliefs (e.g., their sense of whether knowledge is certain and scientific talent innate). Thus, within a single semester, the inexpensive C.R.E.A.T.E. method can shift not just students’ analytical abilities and understanding of scientists as people, but can also positively affect students’ confidence with analysis of primary literature, their insight into the processes of science, and their beliefs about learning.

INTRODUCTION

As scientific information continues to accumulate at a rapid pace, there is a growing sense among science educators that long-established practices need to be reconsidered. Numerous 21st-century science reform documents (American Association for Higher Education, 2000; U.S. Department of Education, 2000; National Research Council [NRC], 2003; Malcom et al., 2005; Alberts, 2005; NRC 2007, 2009; American Association for the Advancement of Science [AAAS], 2010) suggest focusing less on content coverage and more on approaches that reveal science to be an ongoing creative process. Ideally, such a change would help to stem the long-standing attrition of bright students from science majors and, by extension, science research careers (Seymour and Hewitt, 1997; Cech and Kennedy, 2005; DePass and Chubin, 2009). At the same time, student attitudes toward learning (student epistemologies; Schommer 1990, 1993), student self-efficacy (confidence in ability to work effectively in a particular context; Lawson et al., 2007), and student attitudes about science (Osborne, 2003) have been demonstrated to be important factors affecting students’ success in the science classroom. Thus, both changes in how science is taught and consideration of factors influencing students’ ability to learn deserve focus in science education reform efforts.

A variety of new approaches that employ alternatives to lectures, including hands-on classroom activities and
small-group work (Klionsky, 1998; Handelsman et al., 2004; Allen and Tanner, 2005; Knight and Wood, 2005), highlighting controversy to stimulate student engagement (Seethaler, 2005; Campion et al., 2009), student participation in ongoing grant-funded research projects (Hanauer et al., 2006; Call et al., 2007; Lopatto et al., 2008; Clark et al., 2009), case study approaches (Herreid 1994a; Chaplin, 2009), use of the popular press (Strauss, 2005; Hoskins, 2010), and analysis of primary literature (Herreid, 1994b; Janick-Buckner, 1997; Lynd-Balta, 2006; Kozercaki et al., 2006; Hoskins et al., 2007; Hoskins, 2008; Schinske et al., 2008; Yarden, 2009), shift classroom focus from a teacher-centered situation in which students are largely passive, to a student-centered classroom (Freeman et al., 2007; Klymkowski, 2007; Armbruster et al., 2009; Hoskins and Stevens, 2009) more supportive of cognitive activities associated with learning (Bloom and Krathwohl, 1956; Chickering and Gamson, 1987; Zull, 2002).

We have focused on primary literature as a portal into the scientific research process through the C.R.E.A.T.E. (Consider, Read, Elucidate hypotheses, Analyze and interpret data, Think of the next Experiment) method. C.R.E.A.T.E. uses intensive critical analysis of a series of papers generated sequentially from one lab, coupled with email interviews of paper authors via a survey of student-generated questions, to demystify and humanize science. The approach is an iterative method, whereby individual steps of the process (Figure 1 and Table 1) provide students an organized approach to individual journal articles that are “dissected” using a series of novel or adapted pedagogical tools in preparation for intensive class discussion. Papers from a single lab are read in series (Figure 1), allowing students to follow the arc of a research project as it actually progressed. Students are not provided with the full series of papers in advance, nor with the titles, authors, abstracts, or discussions of the papers under consideration. While students could Google the missing information, doing so is ultimately more a hindrance, blunting creative thought, than a help. We encourage students to instead treat the course as a process of discovery. By working with a suite of novel or adapted pedagogical tools to prepare for class, students are empowered to participate actively in the lab-meeting atmosphere of the class sessions, where figures and tables are examined individually, and the logic of the overall study is examined. Previous work has documented precourse versus postcourse shifts in C.R.E.A.T.E. students’ critical thinking and content integration abilities, as well as changes in self-assessed attitudes about science and scientists, as determined by postcourse interviews and the Student Assessed Learning Gains instrument (Hoskins et al., 2007). We report here the results of a survey designed to examine additional aspects of students’ attitudes, beliefs, and self-assessed abilities, comparing responses pre- and post-C.R.E.A.T.E. course.

Students’ beliefs about learning and knowledge affect their ability to learn and their application of metacognitive strategies, including their integration of prior knowledge with the task at hand, studying for understanding rather than superficial recall, and assessing what they do and do not comprehend. Such approaches can significantly facilitate students’ understanding of science (Hartman, 2002; Schraw et al., 2006; Pulmones, 2010). Students’ attitudes about the nature of knowledge, for example, whether knowledge can change over time, and their attitudes about intelligence, for example, whether it is innate and fixed or malleable, comprise a set of epistemological beliefs that affect learning and understanding (Schommer, 1990). Students’ epistemological understandings are typically less sophisticated than those of their professors (Hogan and Maglienti, 2001), and these views can affect study approaches, as well as the extent to which students persist in challenging tasks (Schommer, 1993, 1994; Hofer, 2004). Reasoning ability has also been linked to epistemological beliefs, with students whose epistemologies are more sophisticated showing enhanced skills (Zeineddin and Abd-El-Khalick, 2010).

As the constructivist C.R.E.A.T.E. method uses a number of activities and pedagogical tools (Table 1) designed to increase both student engagement and metacognition, we hypothesized that students’ attitudes toward primary literature, the practice of science, and the nature of learning might change during the semester. We developed a questionnaire aimed at assessing students’ self-assessed views about science, scientists, the research process, and aspects of learning, and administered it on the first and last days of the 14-wk semester. Analysis of student responses indicates that C.R.E.A.T.E. students shifted significantly in their understanding of...
numerous aspects of scientific research, their approach to reading scientific literature, their confidence in their ability to understand science, and, perhaps most interesting, their epistemological beliefs. Because naïve epistemological beliefs can affect students’ study approaches, learning gains, and ability to interpret complex scientific information (Schommer, 1990; Kardash and Scholes, 1996; Pulmones, 2010), shifting student epistemology can be a first step toward developing attitudes toward knowledge and learning more supportive of student success. In this regard, it is notable the changes reported here were achieved during a single semester in a course that essentially costs nothing to implement and does not involve a hands-on laboratory component.

**METHODS**

Participants in the study were students in an upper-level elective at the City College of New York (CCNY). The class met twice weekly for a total of 140 min (2005; three credit hours) or 200 min (2006–2009; four credit hours) per week. Class size averaged 27 students (range: 19–32). Seven iterations of the course are included in this study. Most students were junior or senior biology majors who had completed the course prerequisites at CCNY: a year of introductory biology, and one semester each of genetics and cell/molecular biology. A few students (<10% of each class) were participants in the CCNY post-baccalaureate program. These students had earned degrees in other fields and returned to college to complete premed requirements. In the seven classes represented, 65% of students were female and 61% were African American, Hispanic, or Native American, all groups currently underrepresented at all levels of academic science (National Science Foundation [NSF], 2002, 2008; Atwell, 2004).

**Presurvey/Postsurvey of Student Attitudes and Self-Rated Abilities**

On the first and last days of the semester, students filled out an anonymous Likert-style survey aimed at elucidating their degree of agreement/disagreement with a series of statements. They also answered several open-ended questions on the survey. The survey statements focused on attitudes and beliefs about issues the C.R.E.A.T.E. approach was designed to address, including students’ self-rated ability to understand and analyze primary literature; whether primary literature had influenced their understanding of science; students’ understanding of the scientific research process; and students’ self-rated science reading ability. Confidence in their ability to “think like a scientist,” understanding of “scientists as people,” and sense of whether research science was an appealing career choice. We designed the survey based on our experiences teaching from primary literature in previous classes, focusing on a variety of issues we had determined to be problematic for previous students. Open-ended questions requiring written answers focused on students’ understanding of the activities undertaken by research scientists were set aside for later analysis. This survey was administered in each C.R.E.A.T.E. class (two sections per year in 2005 and 2008 and one section per year in 2006, 2007, and 2009). The 2005 cohort of students was included in a previous analysis of the effects of the C.R.E.A.T.E. class on student critical thinking, content integration, and attitudes toward science/scientists (Hoskins et al., 2007).
Table 2. Seven summary items used on the C.R.E.A.T.E. survey

<table>
<thead>
<tr>
<th>Item</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>I understand it very well.</td>
<td>5</td>
</tr>
<tr>
<td>I have some understanding.</td>
<td>4</td>
</tr>
<tr>
<td>I'm not sure.</td>
<td>3</td>
</tr>
<tr>
<td>I disagree.</td>
<td>2</td>
</tr>
<tr>
<td>I strongly disagree.</td>
<td>1</td>
</tr>
</tbody>
</table>

Surveys were anonymous and coded with numbers known only to the students themselves, to allow alignment of data for within-subject statistical analysis. The survey included 52 statements to which students reacted by marking “I strongly agree,” “I agree,” “I’m not sure,” “I disagree,” or “I strongly disagree” on their survey sheets. Some sample statements were phrased positively (e.g., “I could make a simple diagram that provided an overview of an entire experiment”) and others negatively (e.g., “I do not have a good sense of what motivates people to go into research”).

Three additional propositions were aimed at eliciting students’ overall sense of their ability to read/analyze journal articles, their understanding of the nature of science, and the extent to which primary literature had helped them to understand the nature of science (A, B, C). Question C had four possible responses ranging from “no influence” to “major influence,” and all other questions had five possible responses, phrased in parallel to the question posed (e.g., for question B, on understanding of the nature of science, possible answers ranged from “no understanding” to “understand it very well”). Postcourse, all surveys for which both “pre” and “post” copies were available were scored on a five-point scale, with “strongly agree” = 5 and “strongly disagree” = 1. The additional questions were scored in a parallel way, with scores for C ranging from 1–4 rather than 1–5.

RESULTS

The C.R.E.A.T.E. Survey

The C.R.E.A.T.E. survey included a collection of seven summary items (Table 2) and 52 specific skill and attitude items deemed relevant to the course goals. Following the accrual of 140 cases for the summary items and 155 cases for the skill and attitude items, the data were used to improve the usefulness of the survey. Initial inspection of the 52 skill and attitude items revealed some items were repetitious and others were unrelated to other items (low communalities) or to summary items. These items were set aside, leaving 38 candidates for continued analysis. Of the 38 items, 13 were similar to items used in research on epistemological attitude (Schommer, 1990) and attitude toward science. These are described below (Tables 5, 6, and related text). The epistemological items were a sampling of items developed by Schommer (1990) as part of a much broader investigation of epistemological beliefs. The items here were drawn from across the factors derived from Schommer’s survey, and so include items representing the belief that knowledge is certain (e.g., that different scientists will come to similar conclusions) and that ability is innate (e.g., that scientists were born with a special talent), as well as items assessing attitude toward science (e.g., science is a creative endeavor). Because Schommer’s previous work showed that these epistemological beliefs do not constitute a single scale, we did not include them in exploratory factor analysis, but analyzed them separately. The remaining 25 items were analyzed by means of a principal component analysis (PCA) to explore underlying factors that might aid in the reduction of the 25 variables to a more manageable set. The PCA was performed on 25 skill and attitude items, with a varimax rotation to aid with interpretation. The resulting analysis yielded eight factors accounting for 64% of the variance in the data; however, some variables were “split” across factors, resulting in two factors that were uninterpretable. These were set aside. The remaining six factors, with their member items and factor loadings, are shown in Table 3. The first factor, which we name Decoding Primary Literature, includes items that refer to scientific language and scientific literature, and indicates the respondent’s feelings about reading primary scientific literature. The second factor, Interpreting Data, includes items that have to do with data presented in tables and graphs, as well as with data transformations. The third factor, Active Reading, includes items about diagrams, displays, and method. Visualization, the fourth factor, includes visualizing the method of a study and interpreting graphs. The fifth factor is named Thinking Like a Scientist and includes items about explaining a scientific paper and thinking of experiments. The final factor is called Research in Context and includes items about animal models and controls in experiments.

Pretest–Posttest Differences

Raw scores for the items in each factor were summed, resulting in six pretest scores and six posttest scores for each student respondent. A paired-difference t test was performed on each of the six factors. The results are shown in Table 4. Each pretest–posttest difference is highly significant in the expected direction of posttest gains. The magnitude of the change, estimated as standard deviations in the final column of Table 4, is medium to large (Cohen, 1988). This magnitude of change, as well as the stringent level for significance testing, argues against the presence of a type I error (spurious significant differences).
Table 3. Items from the C.R.E.A.T.E. survey arranged according to a PCA with varimax rotation

<table>
<thead>
<tr>
<th>Factor</th>
<th>Item</th>
<th>Factor loading</th>
<th>Cronbach’s alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Decoding Primary Literature</td>
<td>The scientific literature is difficult to understand (R).</td>
<td>0.776</td>
</tr>
<tr>
<td></td>
<td></td>
<td>When I see scientific journal articles, it looks like a foreign language to me (R).</td>
<td>0.593</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I am not intimidated by the scientific language in journal articles.</td>
<td>0.558</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I am confident in my ability to critically review scientific literature.</td>
<td>0.500</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I am comfortable defending my ideas about experiments.</td>
<td>0.328</td>
</tr>
<tr>
<td>2</td>
<td>Interpreting Data</td>
<td>It is easy for me to transform data, like converting numbers from a table to percents.</td>
<td>0.796</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If I see data in a table, it is easy for me to understand what it means.</td>
<td>0.680</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If I am shown data (graphs, tables, charts), I am confident that I can figure out what it means.</td>
<td>0.622</td>
</tr>
<tr>
<td></td>
<td></td>
<td>It is easy for me to relate the results of a single experiment to the big picture.</td>
<td>0.352</td>
</tr>
<tr>
<td>3</td>
<td>Active Reading</td>
<td>I could make a simple diagram that provides an overview of an entire experiment.</td>
<td>0.763</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If I am assigned to read a scientific paper, I typically look at the methods section to understand how the data were collected.</td>
<td>0.584</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I do not know how to design a good experiment (R).</td>
<td>0.522</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The way that you display your data can affect whether or not people believe it.</td>
<td>0.345</td>
</tr>
<tr>
<td>4</td>
<td>Visualization</td>
<td>When I read scientific information, I usually look carefully at the associated figures and tables.</td>
<td>0.694</td>
</tr>
<tr>
<td></td>
<td></td>
<td>When I read scientific material it is easy for me to visualize the experiments that were done.</td>
<td>0.649</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If I look at data presented in a paper, I can visualize the method that produced the data.</td>
<td>0.592</td>
</tr>
<tr>
<td></td>
<td></td>
<td>When I read a paper, I have a clear sense of what physically went on in a lab to produce the results and information I am reading.</td>
<td>0.584</td>
</tr>
<tr>
<td>5</td>
<td>Thinking Like a Scientist</td>
<td>After I read a scientific paper, I don’t think I could explain it to somebody else (R).</td>
<td>0.735</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I am confident I could read a scientific paper and explain it to another person.</td>
<td>0.655</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I enjoy thinking of additional experiments when I read scientific papers.</td>
<td>0.394</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I accept the information about science presented in newspaper articles without challenging it (R).</td>
<td>0.231</td>
</tr>
<tr>
<td>6</td>
<td>Research in Context</td>
<td>Experiments in “model organisms” like the fruit fly have led to important advances in understanding human biology.</td>
<td>0.774</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Progress in curing diseases has been made as a result of experiments on lower organisms like worms and flies.</td>
<td>0.597</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I understand why experiments have controls.</td>
<td>0.540</td>
</tr>
</tbody>
</table>

* Items followed by an (R) are reverse-scored. Cronbach’s alpha, an index of inter-item consistency, is also shown.

**Epistemological Beliefs**

Table 5 shows 13 items related either to Schommer’s constructs of certain knowledge and innate ability or to a general attitude toward science. It would be expected that students with an insightful attitude toward science would believe that knowledge is not certain and unchangeable; that scientific ability is not fixed and innate, and that science is a creative and collaborative endeavor. To explore the change in these variables from pre- to posttest, the items under the certain knowledge heading were summed (Cronbach’s alpha = 0.66 for this scale). Similarly, the two items for innate knowledge were summed. The attitude items were examined individually. The result of paired-difference t-tests for pre- versus postcourse data are shown in Table 6. There were significant positive gains on all the variables. The possibility of the presence of a type I error is reduced by the stringent level for significance and the moderate effect sizes.

**Relationships to Summary Item of Reading Confidence**

In an effort to synthesize the findings, we chose to ask whether any of the components of the C.R.E.A.T.E. survey identified so far relate to the student participant’s overall view of his or
**Table 5.** Items from the C.R.E.A.T.E. survey that measure epistemological beliefs

<table>
<thead>
<tr>
<th>Knowledge is certain.</th>
<th>If two different groups of scientists study the same question, they will come to similar conclusions. (R)</th>
<th>The data from a scientific experiment can only be interpreted in one way. (R)</th>
<th>Because scientific papers have been critically reviewed before being published, it is unlikely that there will be flaws in scientific papers. (R)</th>
<th>Because all scientific papers are reviewed by other scientists before they are published, the information in the papers must be true. (R)</th>
<th>Sometimes published papers must be reinterpreted when new data emerge years later.</th>
<th>Results that do not fit into the established theory are probably wrong. (R)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ability is innate.</td>
<td>I think professionals carrying out scientific research were probably straight-A students as undergrads. (R)</td>
<td>You must have a special talent in order to do scientific research. (R)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attitude toward science.</td>
<td>Science is a creative endeavor.</td>
<td>I have a good sense of what research scientists are like as people.</td>
<td>I do not have a good sense of what motivates people to go into research. (R)</td>
<td>Scientists usually know what the outcome of their experiments will be. (R)</td>
<td>Collaboration is an important aspect of scientific experimentation.</td>
<td></td>
</tr>
</tbody>
</table>

| Table 5. Items from the C.R.E.A.T.E. survey that measure epistemological beliefsa |
|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|
| Knowledge is certain. | If two different groups of scientists study the same question, they will come to similar conclusions. (R) | The data from a scientific experiment can only be interpreted in one way. (R) | Because scientific papers have been critically reviewed before being published, it is unlikely that there will be flaws in scientific papers. (R) | Because all scientific papers are reviewed by other scientists before they are published, the information in the papers must be true. (R) | Sometimes published papers must be reinterpreted when new data emerge years later. | Results that do not fit into the established theory are probably wrong. (R) |
| Ability is innate. | I think professionals carrying out scientific research were probably straight-A students as undergrads. (R) | You must have a special talent in order to do scientific research. (R) |
| Attitude toward science. | Science is a creative endeavor. | I have a good sense of what research scientists are like as people. | I do not have a good sense of what motivates people to go into research. (R) | Scientists usually know what the outcome of their experiments will be. (R) | Collaboration is an important aspect of scientific experimentation. |

DISCUSSION

**Experiences That Shift Students’ Attitudes about Science and Their Own Scientific Abilities**

The one-semester C.R.E.A.T.E. course enhanced students’ confidence in their ability to read, understand, and explain science, as well as to understand how research is carried out (Tables 3 and 4). Students changed significantly on summary variables that assessed self-rated ability to design experiments, visualize methods based on data, visualize lab activities based on the written account in the journal article, manipulate data, relate results of individual experiments to “the big picture,” critically review data, read science with appropriate skepticism, and explain results to others. These topic areas touch on a wide range of activities undertaken by working researchers.

Our survey findings are consistent with a model that predicts that students’ overall confidence in their ability to read and analyze journal articles is based largely on their sense that they can: 1) Decode Primary Literature (Table 3, factor 1) and 2) Think Like a Scientist (Table 3, factor 5). The gains in students’ confidence in these abilities are likely related to the design of the C.R.E.A.T.E. course. Class discussion and homework assignments challenge students to work through the experiments and interpret the findings of published studies as if they had done the work themselves. During the semester, students carry out numerous activities typical of working scientists. As such, the approach may promote more epistemological engagement than does a standard class, thus engendering changes in students’ attitudes and beliefs about science.

Overall, the approaches typical of the C.R.E.A.T.E. classroom are likely to encourage students to employ metacognitive strategies that help them interpret complex information (Hartman, 2002; Nordell, 2009). Gaining deeper

| Table 6. The results of paired-difference t tests for items (certain knowledge, innate ability, and attitude toward science) in Table 5 |
|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|
| Item                               | Pretest mean (SD) | Posttest mean (SD) | Statistical significance | Mean difference/SD of the differencea |
| Certain knowledge                   | 19.7 (2.2)        | 20.7 (2.7)        | p < 0.001                | 0.40                               |
| Innate ability                      | 7.5 (1.7)         | 8.1 (1.5)         | p < 0.001                | 0.36                               |
| Creativity                          | 4.1 (0.85)        | 4.4 (0.73)        | p < 0.001                | 0.30                               |
| Sense of scientists                 | 3.1 (0.93)        | 3.8 (0.77)        | p < 0.001                | 0.30                               |
| Sense of motives                    | 3.6 (0.95)        | 4.0 (1.0)         | p < 0.001                | 0.31                               |
| Known outcomes                      | 4.0 (0.82)        | 4.3 (0.81)        | p < 0.001                | 0.30                               |
| Collaboration                       | 4.4 (0.73)        | 4.6 (.66)         | p < 0.006                | 0.22                               |

aEstimate of the magnitude of the effect.
understanding of challenging material may in turn help students overcome widely held misconceptions about research science, for example, that experiments are done to demonstrate concepts already known, and therefore research is not a creative activity, and that researchers are simply fact-gatherers (Sandoval, 2003). Students’ repeated opportunities to design and evaluate experiments and models throughout the C.R.E.A.T.E. semester (Figure 1 and Table 1) may contribute to the shifts we noted in their views about the creativity of science (Tables 5 and 6). Developing their own questions for paper authors (e.g., “Do you have to be a straight-A student to become a researcher?” “What would be your ‘dream discovery?’”) encourages students to think beyond the data of the papers and consider the overall process of becoming and being a scientist.

Currently, undergraduate research experiences (UREs) are considered to be one of the most important mechanisms for stimulating students’ interest in science careers. The effects of UREs have been investigated using surveys of students’ experiences (Rauckhorst et al., 2001; Lopatto, 2004a,b, 2007, 2009; Russell, 2006; Russell et al., 2007) and extended interviews (Hunter et al., 2007). These studies have reported high student enthusiasm for UREs, as well as major benefits for multiple aspects of students’ understanding of science, their hands-on research skills, and their attitudes toward research careers. In a longitudinal ethnographic study of the effects of UREs on students at four liberal arts institutions, participants noted in interviews that UREs increased their research confidence and sense of “feeling like a scientist” (Hunter et al., 2007). Students who participate in science UREs are likely to already be interested in scientific research, and UREs clearly reinforce the aspirations of these students. In this respect, C.R.E.A.T.E. may reach a broader group of students, including many who have not previously considered careers in science.

The concept of “nature of science” includes, for many science educators, the ideas that scientists build understanding on observations of nature, that explanations and understanding can change over time, and that creativity comes into play throughout the research process (Lederman, 1992; Karakas, 2009). While there is general agreement that students need to understand the nature and processes of science, or “science as a way of knowing” (AAAS, 1993, p. 2; see also AAAS 1989, 2010, NRC 2009), it is less clear how to accomplish this goal. Teaching approaches focused on inquiry have been suggested as a way to build student understanding of the nature of science (Auills and Shore, 2008; Shore et al., 2008) and enhance learning (Quitadamo et al., 2008). “Inquiry” alone, however, may not be sufficient to shift students’ concepts of the nature of science (Lederman et al., 1998; Sandoval, 2003; Schwartz et al., 2004) or to encourage students to use metacognitive approaches when studying science (Butler et al., 2008).

Our study finds that, although the C.R.E.A.T.E. course does not include a laboratory component or independent research projects, students nevertheless report substantial changes in attitudes and beliefs about science during the semester. Being challenged to devise their own research questions, analyze and interpret data, design experiments, and carry out peer review of studies devised by other students may stimulate C.R.E.A.T.E. participants to examine their personal beliefs about science. In addition, student interview data suggest the email survey of authors plays a role in shifting students’ understanding of “who does science, and why?” (see Tables 1 and S1 in Hoskins et al., 2007). In this context, it is notable that students participating in a novel Deconstructing Scientific Research course, which focuses on intensive analysis of an individual research seminar and also lacks a hands-on component, showed large gains in multiple categories addressed by the Survey of Undergraduate Research Experiences instrument (Lopatto, 2004a, 2007, 2009), including the ability to “understand how knowledge is constructed” (Clark et al., 2009).

Epistemologies, Learning, and the Nature of Science

Our survey also addressed aspects of students’ epistemological beliefs. Schommer and colleagues (Schommer, 1990; Schommer et al., 1992) have identified epistemological beliefs that moderate learning in a variety of intellectual domains. For example, the beliefs that knowledge is certain, that authority should be trusted, that learning is quick and simple, and that intellectual talent is innate can interfere with striving to learn. The C.R.E.A.T.E. program, by uncovering the process of scientific thinking and by providing contact between students and professionals, may influence these epistemological beliefs in a beneficial way. We found substantial changes during the semester in students’ views of “scientists”; moderate shifts in students’ sense of whether knowledge is certain and ability is innate, the creativity of science, or understanding of motives that drive scientists; and a small shift in students’ views of science as a collaborative activity (Tables 5 and 6).

Undergraduates’ epistemological beliefs shift during the college years from a sense that knowledge is certain, typical of freshmen, to a more nuanced view of the relative nature of knowledge, held by seniors (Ferry, 1970), and a longitudinal study suggests such views continue to change postcollege (Baxter Magolda, 2004). Epistemological beliefs change slowly during the college years (Ferry, 1970), and only a minority of students achieve mature epistemological understanding by senior year (Baxter Magolda, 1992). For both high school (Schommer, 1993) and college (Schommer, 1990, 1993; Hofer, 2000, 2004) students, the sophistication of their epistemological beliefs correlates with their reading comprehension and academic performance, with naïve beliefs linked to lesser achievement. Epistemological beliefs also affect student metacognition (Hartman, 2002; Schommer-Aikins, 2002; Hofer, 2004), ability to integrate information (Schommer, 1993), and persistence when confronted with a challenging task (Dweck and Leggett, 1988). The interrelationships among personal epistemologies, metacognition, and learning are complex, but there is general agreement that naïve epistemologies may interfere with learning (Hofer, 2004; Pulmones, 2010). Overall, students with naïve epistemologies employ fewer of the metacognitive strategies (e.g., setting goals, monitoring progress, self-questioning, and connecting new information to broader concepts) that support self-directed learning (Zimmerman, 1990; Hartman, 2002; Pieschel et al., 2008; Stromso et al., 2008).

The C.R.E.A.T.E. method’s combination of epistemically challenging approaches applied in an authentic context may underlie the changes we saw in students’ epistemological beliefs. Several aspects of the C.R.E.A.T.E. approach present students with novel cognitive challenges and associated epistemic load. C.R.E.A.T.E. students employ visualization
when sketching cartoons that fill the gap between the methods section and the charts, graphs, blots, and/or photomicrographs presented. Integrating verbal information with visual information promotes integration of different modalities. Such integrative thinking, reinforced by C.R.E.A.T.E.’s repeated use of concept maps, both as a tool for review and a way to organize papers’ central themes, can facilitate learning (Novak, 1991; Van Meter and Garner, 2005; Schwartz and Heiser, 2006). Class discussion often focuses on a point of controversy, which can both increase student engagement (Bell and Linn, 2002) and stimulate students to “do the real intellectual ‘work’ of synthesizing ideas across subdomains” (Seethaler, 2005, p. 273). C.R.E.A.T.E.’s narrow focus on a few papers may encourage students to work toward deep rather than superficial understanding (Schwartz et al., 2009) as they engage in cognitively stimulating activities corresponding to upper levels of Bloom’s taxonomy (e.g., analysis, synthesis, evaluation: levels 4–6; Bloom and Krathwohl, 1956). Extended discussions involving scientific argumentation are rare in lecture-dominated classrooms (Osborne, 2010), but can be of substantial benefit, especially when students feel free to develop their understanding through discussion and to speculate aloud as they do so (Sawyer, 2006). C.R.E.A.T.E.’s “grant panel” activities encourage student reflection on the research process beyond the details of individual journal articles.

UREs might also be expected to have a strong effect on students’ epistemological beliefs. This has been seen in some cases (Rauckhorst et al., 2001; Lopatto, 2004b; both studies include both science and nonscience URE participants), but in other studies epistemological beliefs appeared not to shift significantly during the URE, either as reported by student participants or their faculty mentors (Hunter et al., 2007). A recent meta-analysis of independent research experiences in science (Sadler et al., 2010) suggests that supplementing research experiences with specific additional activities, such as keeping a journal of reflections on the research experience (Rauckhorst et al., 2001, college students; Schwartz et al., 2004, high school teachers) or interacting with peers also involved in research apprenticeships (Grindstaff and Richmond, 2008; high school students), can expand gains made in UREs and enhance understanding of the nature of science. These researchers further note that developing a deeper understanding of the nature of science will probably require instructional approaches that ensure undergraduates’ participation in developing hypotheses and analyzing data, both considered “epistemically demanding practices” (Sadler and McKinney, 2010, p. 48). Other investigators have suggested that epistemological beliefs are more likely to change if students are trained to think critically in a context that encourages metacognition and includes controversy (Valanides and Angeli, 2005). In studies of high school students (Bell et al., 2003) and undergraduates (Ryder et al., 1999), changes in scientific thinking were seen in students who participated in projects that demanded substantial epistemic engagement. Conversely, classrooms lacking in authentic scientific inquiry activities can reinforce naïve epistemological beliefs, for example, that scientific logic is simple and conclusions certain (Chinn and Malhotra, 2002). Students’ ability to carry out research projects may be constrained by such beliefs (Ryder and Leach, 1999).

We consider it likely that the shifts we see in epistemological beliefs of C.R.E.A.T.E. students are attributable to students’ experiences in the C.R.E.A.T.E. course. We did not measure presemester/postsemester epistemological beliefs in an independent control group of students who did not take the course, as no such control group was available. We are not, however, aware of any studies that show that increased sophistication of epistemological beliefs results from mere maturation or passage of time during a single semester. The experiences of science students in UREs may provide some insight into the malleability of students’ epistemological beliefs. Science students’ UREs would be expected to support or enhance any shifts in their epistemological beliefs that occurred “maturationally” during an academic semester. Thus, the finding that epistemological beliefs tend to remain stable in science URE participants interviewed repeatedly over several years (Hunter et al., 2007), suggests that the epistemological beliefs of undergraduate science students do not shift rapidly. We feel it is likely that the postsemester versus presemester changes we document were brought about by students’ experiences in the semester-long C.R.E.A.T.E. course.

Although the C.R.E.A.T.E. approach is not unique in focusing on primary literature, it is unusual in its combination of intensive analysis of a series of related publications with an email survey of their authors, and its concentration in the classroom on discussion and analysis aimed at simultaneously decoding the figures and tables, modeling the research process, and humanizing the scientists behind the papers. The C.R.E.A.T.E. teaching method encourages students to engage in conversations, debates, and creative thinking, which involve cognitive challenges that can help develop understanding of complex material (Driver et al., 2000; Marbach-Ad and Sokolove, 2000; Bell and Linn, 2002; Seethaler, 2005; Campion et al., 2009) and at the same time encourage creative approaches to such material (DeHaan, 2009). Overall, our findings indicate that the C.R.E.A.T.E. method increases students’ confidence in their ability to read and understand primary literature, improves their self-assessed understanding of the nature and processes of science, and encourages their development of more sophisticated epistemological beliefs. We suggest that complementing existing curricula with inexpensive C.R.E.A.T.E.-style courses could be an effective way to help students develop deeper insight into the nature and practices of science. Finally, students who recognize early in their college years that science is creative and open-ended might be more likely to take advantage of the UREs that can stimulate and reinforce interest in science research careers.

ACKNOWLEDGMENTS

Many thanks to CCNY students for participating in the C.R.E.A.T.E. course and associated assessments. The C.R.E.A.T.E. project is approved by CUNY IRB, protocol H-0633. We thank the NSF for support. S.G.H. also thanks the Helen Riaboff Whiteley Center at the Friday Harbor Laboratories, Friday Harbor, Washington, for academic support during a recent sabbatical, and Ross Nehm for discussion of survey design. This material is based upon work supported by the NSF CCLI/TUES program under Grants 0311117, 0618536, and 1021443. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.
REFERENCES

Allen D, Tanner K (2005). Infusing active learning into the large-
enrollment biology class: seven strategies, from the simple to com-
American Association for Higher Education (2000). Targeting Curric-
ular Change: Reform in Undergraduate Education in Science, Math,
Engineering, and Technology, Washington, DC: American Associa-
tion for Higher Education.
American Association for the Advancement of Science (AAAS)
Literacy Goals in Science, Mathematics, and Technology,
Washington, DC: AAAS. www.project2061.org/publications/
University Press.
visionandchange.org/files/2010/05/VC_report.pdf (accessed 23 July
2011).
and student-centered pedagogy improve student attitudes and per-
Atwell R (2004). The long road ahead: barriers to minority partic-
ipation persist. In: Reflections on 20 Years of Minorities in Higher
Education and the ACE Annual Status Report, Washington, DC:
American Council on Education, Center for Advancement of Racial
and Ethnic Equity. www.acenet.edu/Content/NavigationMenu/
ProgramsServices/CAREE/2004_reflections_msr.pdf (accessed 23
July 2011).
Foundations for Research as a Curricular Imperative, New York:
Lawrence Erlbaum.
Baxter Magolda MB (1992). Knowing and Reasoning in College:
Gender-Related Patterns in Students’ Intellectual Development, San
Baxter Magolda MB (2004). Evolution of a constructivist conceptual-
a science apprenticeship program on high school students’ un-
derstandings of the nature of science and scientific inquiry. J Res
Sci Teach 40, 487–509.
Bell P, Linn M (2002). Beliefs about science: how does science in-
struction contribute? In: Personal Epistemology: The Psychology of
Beliefs about Knowledge and Knowing, ed. B. Hofer and P. Pintrich,
Bloom BS, Krathwohl DR (1956). Taxonomy of Educational Objectives:
The Classification of Educational Goals, by a Committee of
College and University Examiners, Handbook I: Cognitive Domain,
New York: Longman.
authentic inquiry in the sciences: challenges faced in redefining uni-
versity students’ scientific epistemology. In: Inquiry in Education,
vol. II., ed. B. Shore, M. Aulls, and M. Delcourt, New York: Rout-
ledge, 301–324.
Call GB, et al. (2007). Genomewide clonal analysis of lethal mutations
in the Drosophila melanogaster eye: comparison of the X chromosome
and autosomes. Genetics 177, 689–97.
Campion N, Martins D, Wilhelm A (2009). Contradictions and pre-
dictions: two sources of uncertainty that raise the cognitive interest
Chaplin S (2009). Assessment of the impact of case studies on student
learning gains in an introductory biology course. J Coll Sci Teach 39,
72–79.
Chinn C, Malhotra B (2002). Epistemologically authentic inquiry in
schools: a theoretical framework for evaluating inquiry tasks. Sci
Educ 86, 175–218.
Clark IE, Romero-Calderón R, Olson JM, Jaworski L, Lopatto D,
and scalable pedagogical tool to provide evidence-based science in-
Cohen J (1988). Statistical power analysis for the behavioral sciences,
DeHaan R (2009). Teaching creativity and inventive problem solving
Encourage Minorities to Pursue Research Careers: Summary of a
Driver R, Newton P, Osborne J (2000). Establishing the norms of
Dweck CS, Leggett EL (1988). A social-cognitive approach to moti-
vation and personality. Psychol Rev 95, 256–273.
Freeman S, O’Connor E, Parks JW, Cunningham M, Hurley D, Haak
D, Dirks C, Wenderoth MP (2007). Prescribed active learning in-
creases performance in introductory biology. CBE Life Sci Educ 6,
132–139.
Grindstaff K, Richmond G (2008). Learners’ perceptions of the role
of peers in a research experience: implications for the apprenticeship
process, scientific inquiry, and collaborative work. J Res Sci Teach 45,
251–271.
Hanauer D, Jacobs-Sera D, Pedulla M, Cresawn S, Hendrix R, Hatfull
522.
Hartman H (2002). Developing students’ metacognitive knowledge
and skills. In: Metacognition in Learning and Instruction: Theory,
Research and Practice, ed. H. Hartman, Dordrecht, The Netherlands:
Herreid CF (1994a). Case studies in science—a novel method of sci-
Herreid CF (1994b). Journal articles as case studies: The New Eng-
355.
Hofer BK (2000). Dimensionality and disciplinary differences in per-
Hofer BK (2004). Exploring the dimensions of personal epistemology
in differing classroom contexts: student interpretations during the
Hogan K, Maglienti M (2001). Comparing the epistemological under-
pinnings of students’ and scientists’ reasoning about conclusions. J
Res Sci Teach 38, 663–687.
Hoskins S (2008). Using a paradigm shift to teach neurobiology and
Neurosci Educ 6, A40–A52.
Hoskins S (2010). “But if it’s in the newspaper, doesn’t that mean
it’s true?” Developing critical reading and analysis skills by evalu-
ating newspaper science with C.R.E.A.T.E. Amer Biol Teach 72, 415–
420.
Hoskins SG, Stevens LM (2009). Learning our L.I.M.I.T.S.: less is more


Schommer M (1994). An emerging conceptualization of epistemological beliefs and their role in learning. In: Beliefs about Text and


Exploring Undergraduates’ Understanding of Photosynthesis Using Diagnostic Question Clusters

Joyce M. Parker,*† Charles W. Anderson,‡ Merle Heidemann,† John Merrill,§ Brett Merritt,‡ Gail Richmond,‡ and Mark Urban-Lurain||

*Department of Geological Sciences, ‡Department of Teacher Education, §Biological Sciences Program, and ||College of Engineering, Michigan State University, East Lansing, MI 48824

We present a diagnostic question cluster (DQC) that assesses undergraduates’ thinking about photosynthesis. This assessment tool is not designed to identify individual misconceptions. Rather, it is focused on students’ abilities to apply basic concepts about photosynthesis by reasoning with a coordinated set of practices based on a few scientific principles: conservation of matter, conservation of energy, and the hierarchical nature of biological systems. Data on students’ responses to the cluster items and uses of some of the questions in multiple-choice, multiple-true/false, and essay formats are compared. A cross-over study indicates that the multiple-true/false format shows promise as a machine-gradable format that identifies students who have a mixture of accurate and inaccurate ideas. In addition, interviews with students about their choices on three multiple-choice questions reveal the fragility of students’ understanding. Collectively, the data show that many undergraduates lack both a basic understanding of the role of photosynthesis in plant metabolism and the ability to reason with scientific principles when learning new content. Implications for instruction are discussed.

INTRODUCTION

The goal of the work presented here was to develop a diagnostic question cluster (DQC) that would yield information on undergraduates’ thinking about photosynthesis to inform improvements in instruction and assessment. This assessment tool is diagnostic in the sense that it identifies patterns across students’ responses to questions, revealing root problems that can be the focus of instructional change.

Our work indicates that in order to apply basic concepts about photosynthesis, students need to be able to engage in a coordinated set of practices based on a few scientific principles: conservation of matter, conservation of energy, and the hierarchical nature of biological systems. We work with clusters of questions, rather than individual questions, to assess students’ abilities to do the coordinated practices and to see whether students’ abilities to apply concepts are context specific.

BACKGROUND

Misconceptions about Photosynthesis

Misconceptions about photosynthesis are well documented (e.g., Eisen and Stavy, 1988; Amir and Tamir, 1994; Hazel and Prosser, 1994; Marmaroti and Galanopoulou, 2006; Yenilmez and Tekkaya, 2006; Köse, 2008). These are pervasive and persist throughout schooling, from primary to postsecondary education. Some of these misconceptions arise from direct experiences students have had observing plants. For example, the idea that plants obtain all of their nutrients from the soil matches everyday experience with plants, in which the only visible inputs are through the roots (Eisen and Stavy, 1988; Marmaroti and Galanopoulou, 2006; Köse, 2008). Other misconceptions are perpetuated...
Three principles we have found useful are conservation of matter (Wilson et al., 2006), conservation of energy, and the hierarchical nature of biological systems. All three of these principles are identified in the Vision and Change in Undergraduate Biology Education report (American Association for the Advancement of Science [AAAS], 2010) as being “core concepts” in biology education. These principles encompass such statements as:

- During chemical reactions, intramolecular bonds are broken and atoms are rearranged to form molecules of new substances as new bonds form. No atoms are lost in the process.
- Energy is used to break bonds, while energy is released when bonds form.
- Biological systems are nested in scale, and the properties and functions of a particular scale emerge from the properties and functions of smaller scales.

Principled reasoning also involves using a coordinated set of practices related to the scientific principles. For photosynthesis, we have found three practices to be important. We present the content associated with photosynthesis organized around these three practices in Supplemental Material B.

The practice of tracing matter includes:

- identifying the matter that changes, that is, the inputs and outputs of a system or the reactants and products of a reaction or set of reactions;
- distinguishing matter from energy;
- tracing atoms; and
- conserving matter.

In photosynthesis, tracing matter includes knowing the overall reaction \(6\text{CO}_2 + 12\text{H}_2\text{O} \rightarrow \text{C}_6\text{H}_{12}\text{O}_6 + 6\text{H}_2\text{O} + 6\text{O}_2\) and tracing individual elements through the process to see, for example, that elemental oxygen produced does not come directly from carbon dioxide, as shown by the color-coding of oxygen in this reaction.

The practice of tracing energy includes:

- identifying the energy that is transformed or transferred and the forms of energy involved;
- describing the nature of the transformations or transfers, conserving energy, and
- identifying processes that transfer or transform information.

The energy transformations of photosynthesis include transforming sunlight to chemical potential energy in NADPH. That chemical potential energy is transferred from a proton gradient to chemical potential energy in ATP and finally to chemical potential energy in fixed carbon. This establishes that ATP production is not the end point of photosynthesis.

The practice of organizing systems and identifying scale includes:

- Knowing the structure of the systems in which the relevant processes are taking place and how they facilitate the function.
• Selecting the appropriate level/scale in which to reason. In biological systems, the explanations for, or mechanisms of, phenomena apparent at one scale often lie at a different scale. For example, a plant such as a maple tree (at the human scale) gains mass as it grows through the molecular/subcellular process of photosynthesis.

The need for the last principle is perhaps not as obvious as for the first two. However, problems of scale plague much science instruction, impacting discussions of large amounts of time (geology, biology, astronomy), large distances (astronomy), small amounts of time (physics, chemistry, biology), and small sizes (physics, chemistry, biology). In a study of high school students’ understanding of the cell, cellular structures, and processes, Flores et al. (2003) showed that many problems arise because students fail to distinguish between processes that happen at the organismal level versus the organism or cellular level. An example of this type of confusion can be seen when students confuse respiration and digestion (Songer and Mintzes, 1994). Ben-Zvi and Orion (2005) define understanding scale as one of several key components of a systems approach to learning science. They point out that this approach gives students a framework for addressing many topics. We focus on these three fundamental principles, which apply to many topics.

METHODS

Question Development

We developed the cluster of questions used in this study by asking open-ended versions of these questions to undergraduates in large introductory biology classes or smaller upper-level courses. Common inaccurate responses were noted, and these were used to develop distractors for diagnostic multiple-choice questions (Tregust, 1988; Sadler, 1998). A multi-departmental panel of biologists and science teacher educators reviewed the questions for content validity. Construct validity was checked by administering the multiple-choice items to groups of students and asking them to explain their answer choices in writing or through interviews. All revised multiple-choice items were administered postinstruction on standard course exams in an introductory biology course with enrollments of 263–449 students from 2004 to 2009 at a large midwestern university.

We asked two levels of questions. The lower-level questions asked students to identify the inputs and outputs of the light reactions and Calvin cycle or to trace elements or energy transformations through these reactions. These questions did not ask students to carry this information across scales. They asked directly about matter and energy transformations at the cellular or subcellular levels—the scales that are usually emphasized during instruction in an introductory cell biology course. In Bloom’s taxonomy, these would be classified as comprehension questions (Bloom, 1956).

The higher-level questions asked students to apply (Bloom, 1956) what they know about the matter and energy transformations of photosynthesis to explain phenomena in plants. Thus, for these questions, students needed to understand how the whole organism used the cellular process of photosynthesis.

Table 1. Questions as they appeared on various versions of the assessment

<table>
<thead>
<tr>
<th>Question/test</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maple tree</td>
<td>MC</td>
<td>Essay</td>
<td>MC</td>
<td>Essay</td>
</tr>
<tr>
<td>Corn</td>
<td>Essay</td>
<td>MC</td>
<td>Essay</td>
<td>Corn</td>
</tr>
<tr>
<td>Euglena</td>
<td>MC</td>
<td>MC</td>
<td>MT/F</td>
<td>MT/F</td>
</tr>
<tr>
<td>Geranium root</td>
<td>MT/F</td>
<td>MT/F</td>
<td>MC</td>
<td>MC</td>
</tr>
</tbody>
</table>

*MC: multiple choice; MT/F: multiple true/false; Essay: constructed response.

Question Format Comparison

To compare what can be learned from different question formats, we did cross-over experiments comparing multiple-choice with multiple-true/false format and multiple-choice with essay format. (In multiple-true/false format, distracters are presented as individual statements and students indicate whether each one is true or false without knowing how many are true.) Four forms of an exam were generated with different formats of four questions, as seen in Table 1. Students were randomly given one of the four test versions.

The maple tree and corn questions asked students about the source of mass in growing plants. In multiple-choice format, they had the same foils in the same order. The Euglena and geranium root questions asked students about sources of ATP for cells in photosynthetic organisms. The foils were not the same. (For the specific questions, see Supplemental Material A and Results.) The order of questions on all exams was: multiple-choice version of maple tree or corn question, Euglena question, geranium root question, and essay version of maple tree or corn question.

Essays were about the source of mass gain in growing plants (maple tree and corn) and were scored as correct if students mentioned photosynthesis and carbon dioxide. Inaccurate processes and inputs to photosynthesis were noted separately.

Interviews

In a different semester, we conducted interviews with student volunteers in order to gain richer insight into students’ understanding of photosynthesis. A month after taking their hourly exam, students were asked about three of the exam questions. They were asked to explain mass gain in corn plants and radish seeds growing in light and the energy transformations in Euglena growing in light, in that order. Volunteers were sorted into three categories: those answering both mass-gain questions correctly on the exam, those with a mix of correct and incorrect answers, and those with no correct responses. Students were randomly chosen from each category. In total, 14 interviews were performed. During the interviews, the students were shown the stem (question without distracters) to the “radish seeds in light” question (question 7 in Supplemental Material A; Ebert-May et al. 2003) and asked to explain the mass gain. They were then shown the distracters one by one and asked to explain which they would choose (or not choose) and why; the process was repeated with the corn and Euglena questions.
J. M. Parker et al.

Table 2. Demographics of students (n = 333)*

<table>
<thead>
<tr>
<th>Gender</th>
<th>Female</th>
<th>61.9%</th>
<th>Male</th>
<th>38.1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class standing</td>
<td>1</td>
<td>11.1%</td>
<td>2</td>
<td>59.8%</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>21.0%</td>
<td>4</td>
<td>7.8%</td>
</tr>
<tr>
<td>Post-BA, second degree</td>
<td>0.3%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethnicity</td>
<td>Caucasian</td>
<td>82.3%</td>
<td>American Indian</td>
<td>0.3%</td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td>5.4%</td>
<td>Hispanic</td>
<td>3.0%</td>
</tr>
<tr>
<td></td>
<td>Asian</td>
<td>4.5%</td>
<td>Other or not reported</td>
<td>4.5%</td>
</tr>
<tr>
<td>Major</td>
<td>Pre-health</td>
<td>40.5%</td>
<td>Science</td>
<td>18.9%</td>
</tr>
<tr>
<td></td>
<td>Engineering</td>
<td>8.7%</td>
<td>Agriculture</td>
<td>7.8%</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>24.1%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Class standing is based on number of course credits. Multiple majors are included in each category. The pre-health group of majors includes students identifying a major associated with health or medical professions, such as medical technology, pre-nursing, or human biology.

Students

All data were collected in a one-semester course on cell and molecular biology that is one of a two-semester series of courses in introductory biology. Table 2 shows the demographic data for the semester of students who took the exam described in Table 1. This group of students was representative of students in other semesters. This course serves a large number of majors from multiple colleges. Students came from 71 majors. The largest single major was prenursing. One semester of introductory chemistry is a course prerequisite, and sophomores therefore represent the largest population of students.

RESULTS

The DQC and Students’ Responses

Cluster questions, along with data on students’ responses, are shown in Supplemental Material A. All items were administered postinstruction on standard course exams in introductory biology courses with enrollments of 263–449 students. Thirteen questions are presented in multiple-choice format, but can be used in multiple-true/false format (see Comparison of Question Formats below). The stems from most of the questions can be used as prompts for essay questions. Altered stems are proposed for essay versions of the remaining questions.

Questions are categorized according to the practices demanded by the stem. They are presented in order of increasing complexity. The first four questions ask students to trace matter through the process of photosynthesis at the cellular level. Questions 5–9 involve tracing matter across scales, since the questions are posed about whole organisms, but the explanations lie at the cellular level. Of these questions, 5–7 address mass gain in plants, while questions 8 and 9 address mass loss in plants. The latter require that students understand that plants undergo both respiration and photosynthesis. Questions 10–12 ask students about energy sources for plants. In particular, questions 10 and 11 ask about energy sources for cells. However, question 12 asks about cells in different parts of a multicellular plant. These questions require students to trace energy through both photosynthesis and respiration. While the stems of questions can be fairly cleanly categorized in this way, the distracters often encompass inappropriate mixed practices, such as matter–energy conversions or scale mistakes. Thus, as a diagnostic, an item whose stem calls for tracing matter may diagnose incorrect tracing of energy, and so on.

Questions were administered postinstruction on an exam in an introductory biology course at least once between 2004 and 2009. Summary results are shown in Table 3. Results for individual questions are shown in Supplemental Material A.

Comparison of Question Formats

Multiple True/False. Multiple-choice questions are a manageable assessment format for large-enrollment courses, which often require machine scoring. However, students approach them using various test-taking skills to identify the correct answer. These skills, such as length of a distracter, word recognition, and how scientific an answer sounds, can result in inaccurate estimates of students’ understanding. In particular, we were concerned with students who might remember a correct response but still retain inaccurate ideas. Nehm and Schonfeld (2008) describe this as having mixed or heterogeneous understanding. To test the prevalence of this situation, we used a cross-over experimental design. As part

Table 3. Scores on cluster questions categorized by practice(s) required (n = 263–449)*

<table>
<thead>
<tr>
<th>Practices demanded by question</th>
<th>Question numbers</th>
<th>Percent of students answering correctlyb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tracing matter</td>
<td>1–4</td>
<td>34.2–75.9</td>
</tr>
<tr>
<td>Tracing matter and keeping track of scale</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mass gain in plants</td>
<td>5–7</td>
<td>48.6–80.1</td>
</tr>
<tr>
<td>Mass loss in plants</td>
<td>8–9</td>
<td>31.3–56.3</td>
</tr>
<tr>
<td>Tracing energy</td>
<td>10–11</td>
<td>15.4–46.5</td>
</tr>
<tr>
<td>Tracing energy and keeping track of scale</td>
<td>12</td>
<td>31.2</td>
</tr>
</tbody>
</table>

*Varies depending upon the semester in which the questions were asked.

bThe percent correct varies by question, the ranges are for the questions in each subset.
E. The root cell makes ATP by cellular respiration using material absorbed from the soil.

A potted geranium sits in a windowsill absorbing sunlight. How does a root cell (which is not exposed to light) obtain energy to do cellular work such as active transport across its membrane?

A. ATP is made in the leaves via photosynthesis and moved to the root.
B. Sugar is made in the leaves via photosynthesis and moved to the root.
C. The root cell makes sugar using the dark reactions (Calvin cycle) of photosynthesis.
D. The root cell makes ATP by photosynthesis and cellular respiration.
E. The root cell makes ATP by cellular respiration using material absorbed from the soil.

% Students

Figure 1. Percent of students (n = 380) choosing specific multiple-choice distracters (blue) vs. percent of students indicating statement is true (red) in multiple-true/false version of same question. B is the correct answer.

of a standard exam (2009, n = 380), half of the students in an introductory biology course (see Methods) were given a question (the geranium root question) as a multiple-choice question, while the other half of the students were given the same question in a multiple-true/false format in which the distracters were presented as individual statements, and students had to indicate whether each one was true or false without knowing how many were true. Students were given a second question (the Euglena question) in the other format. The results are shown in Figures 1 and 2.

For the geranium root question, students selected the correct answer (B) most frequently, regardless of the format in which the question was delivered. However, in the multiple-true/false format, more than half of the students indicated that the incorrect choice (A) was also true, and at least one-fourth of the students indicated that each of the choices was true, implying that they simultaneously held accurate and inaccurate ideas.

With the Euglena question, students’ mixed ideas about the source of ATP for cellular work are even more apparent. This was a difficult question for students. Regardless of question format, the most popular choice was incorrect—that Euglena use ATP made during photosynthesis to do cellular work. However, in the multiple-true/false format, more than half of the students indicated that four of the five choices were true.

Euglena are single-celled, photosynthetic eukaryotes. How do Euglena obtain energy to do cellular work such as active transport across its membrane?

A. They transport ATP from the chloroplasts.
B. They utilize inorganic nutrients from the surrounding water to make ATP.
C. They use sugars made in the chloroplasts to make ATP.
D. They use the ATP made during photosynthesis.
E. They utilize organic molecules from their surroundings.

% Students

Figure 2. Percent of students (n = 380) choosing specific multiple-choice distracters (blue) vs. percent of students indicating statement is true (red) in multiple-true/false version of same question. C is the correct answer.
**Table 4.** Percent of students correctly answering the multiple-choice and essay versions of the corn and maple tree questions and percent of students including both correct and incorrect mass sources in their essay responses.

<table>
<thead>
<tr>
<th>Question</th>
<th>Multiple choice</th>
<th>Essay</th>
<th>Students with both correct and incorrect sources of mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corn</td>
<td>61.4 ± 8.8%</td>
<td>51.0 ± 7.1%</td>
<td>17 ± 5.4%</td>
</tr>
<tr>
<td></td>
<td>(n = 186)</td>
<td>(n = 190)</td>
<td>(n = 190)</td>
</tr>
<tr>
<td>Maple tree</td>
<td>57.4 ± 9.3%</td>
<td>47.2 ± 7.2%</td>
<td>17 ± 5.5%</td>
</tr>
<tr>
<td></td>
<td>(n = 190)</td>
<td>(n = 186)</td>
<td>(n = 186)</td>
</tr>
</tbody>
</table>

*a* All percentages are shown with 95% confidence intervals.

**Essay.** A crossover experiment was used to compare essay and multiple-choice formats in an introductory biology course (2009, $n = 380$). On an exam, half of the class was given one of two questions about the source of mass in plants (maple tree or corn, see questions 5 and 6 in Supplemental Material A) as an essay question, with the second question being a multiple-choice question. In the multiple-choice version, both questions had the same choices in the same order. Essay responses were scored as correct if students mentioned carbon dioxide as a source of mass for plant growth and named photosynthesis (or the Calvin cycle) as the process, regardless of the presence of incorrect sources of mass or process labels. In a separate round of scoring, students who included both correct and incorrect sources of mass were noted. The results are shown in Table 4.

Fewer students answered the essay version of this question correctly, compared with the multiple-choice version, indicating that they could correctly identify from a list the main contributors to plant mass gain, but could not articulate these ideas on their own. It should be noted that the multiple-choice version of the question appeared first on the exam forms, with the essay version occurring six questions later. However, very few students used the wording of the multiple-choice distracters in their essay, suggesting little copying or transfer.

Like the multiple-true/false questions (see preceding section), the essay responses revealed that students’ understanding included both accurate and inaccurate ideas. Seventeen percent of the students who identified accurate sources for plant mass gain (carbon dioxide and/or water) also mentioned inaccurate sources, such as fertilizer, minerals, or sunlight. (Students who mentioned fertilizer or minerals, but indicated that these were minor contributors, were not counted as having heterogeneous responses.)

**Interviews.** To determine how students interpreted the multiple-choice questions and to see whether their exam responses were indicative of their thinking, 14 student volunteers were interviewed about their midterm exam responses to two questions about the sources of mass when plants (radish seeds and corn) grow and the source of ATP for cellular work in *Euglena*. They were asked to respond in turn to the stem and then the distracters to the radish seed question, the corn question, and the *Euglena* question.

**Mass Gain in Plants—Correct Responders.** Four of the 14 students responded correctly on the exam to both the radish seed and corn questions about the sources of mass gain during plant growth. Excerpts from their interview transcripts are shown in Table 5 (all names are pseudonyms).

<table>
<thead>
<tr>
<th>Question</th>
<th>Excerpt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corn</td>
<td>Burt in response to the radish seed stem: <em>And then how this increase in mass biomass</em> (radish seeds in light) occurred <em>would obviously not be from water</em> so it had to be from something else like <em>some sort of glucose</em> or <em>something like that</em>. It (glucose) is made of C6H12O6 and so it needs the CO2 <em>to make for the carbon and it has water that uses H</em> for <em>the water from the water too</em>.</td>
</tr>
<tr>
<td>Corn</td>
<td>Devin in response to the radish seed stem: <em>All right. With these, with the light and water, it’s so much because during photosynthesis, carbon dioxide is taken in from the outside and then, through a number of processes, it’s released. But the carbon atoms add to the mass of the plants</em>.</td>
</tr>
<tr>
<td>Corn</td>
<td>Devin in response to the radish seed distracters: <em>I know that plants need soil to grow in and that some of the minerals in soil help it to just grow and prosper and do its day-to-day activities, but it’s not a significant contributor to the mass of the plant. I wouldn’t say that this [solar radiation] is a direct contribution, but the plants need solar energy to do photosynthesis, so the solar radiation doesn’t cause the increase in mass. That’s not the materials, but it’s like the light coming down is probably the cause of the... it makes the ATP that is needed for photosynthesis</em>.</td>
</tr>
<tr>
<td>Corn</td>
<td>Wendy in response to the radish seed stem: <em>Okay, we spent like 20 minutes on this in class and so I can answer this without even thinking. The increase in biomass occurred from the intake of carbon dioxide and the expulsion of oxygen</em>.</td>
</tr>
<tr>
<td>Corn</td>
<td>Wendy in response to the stem of corn question: <em>Oh that is easy, it is just because on farms, obviously they use fertilizers which are high in nitrates which can help to increase energy levels and bio mass intake because then they also have the intake from the roots, there is more minerals in the soil, increasing the bio mass, obviously there is carbon dioxide all around</em>.</td>
</tr>
<tr>
<td>Corn</td>
<td>Wendy in response to the distracters for the corn question: <em>It is still the CO2 because that is the main one, the others are just added bonuses to help with different processes because I want to say that there is a cycle</em>.</td>
</tr>
<tr>
<td>Maple tree</td>
<td>Olivia in response to the stem of the radish seed question: <em>I just don’t really remember, so I’m like, I know part of it and I understood... I got it when he was teaching it, but I haven’t reviewed it, so I’m just not quite sure. I just know that there’s something with carbons and that’s why it has more mass... She chose carbon dioxide after seeing the distracters</em>.</td>
</tr>
</tbody>
</table>
able to identify carbon dioxide as the correct response to both questions, she could not describe photosynthesis.

**Mass Gain in Plants—Incorrect Responders.** Six interviewees (Wren, Ruth, Sherman, Phillipa, Winnona, and Ivy) answered both questions about the sources of mass gain in plants incorrectly on the exam. Summaries and excerpts from their interviews are shown in Table 6. Four of the six demonstrated a confused understanding of photosynthesis during the interview. Winnona gave accurate responses in the interview, but explained that on the exam she reasoned from common sense and had forgotten about photosynthesis. Ivy was the only volunteer to choose solar radiation on the exam. In the interview, she gave accurate responses. The interviewer did not ask her about the discrepancy in her answers.

**Mass Gain in Plants—Mixed Responders.** Three students correctly answered the radish seed question, but not the corn question, on the exam. One student correctly answered the corn question, but not the radish seed question, on the exam. They chose the same responses during the interview.

**Energy Sources for Plants.** In the same interviews, students were asked about a third cluster question—the Euglena question. This question asks students about the source of ATP for cellular work in photosynthetic Euglena. Five of the 14 interviewees answered the question correctly on the exam and in the interview. Six said that *Euglena* use ATP made from sunlight both on the exam and in the interview. We thought that they might be interpreting the question to mean that sunlight is the ultimate source of energy for ATP synthesis, a correct statement. However, the interviews showed this not to be the case. Students indicated that ATP was a product of photosynthesis. Some mentioned the electron transport chain producing ATP in response to sunlight.

Three students chose a response in the interview that differed from their exam choice. On the exam, Bob chose the response that *Euglena* use sunlight to make ATP. Before seeing the distractors in the interview, he could only articulate that *Euglena* use photosynthesis. After looking at the distractors, he wavered between that response and the correct answer. Pomina’s responses followed a similar pattern. In the interview, she indicated that ATP made from sunlight and sugars made in the chloroplasts were the best answers. She was unsure whether “photons” or sugars were the most important products of photosynthesis. She eventually chose sugars. Selena chose the correct answer on the exam. In the interview, she rejected this answer, because she remembered that...
organisms could use other types of molecules, such as lipids, to make ATP. Instead, she chose ATP made in the chloroplast, describing the electron transport chain.

Other Insights into Students’ Thinking from Interviews. Interviews are useful for exploring students’ understanding of words, in particular words that have precise meanings in biology, but broader meanings in common usage. Three students associated “solar radiation” with damage, ultraviolet light, or other nondescript negative connotations. An additional student was unsure of the meaning of radiation. Five students had confused or vague ideas of what mineral and organic substances are. Two students associated organic substances with healthy foods or “natural stuff.”

Finally, interviews yield clues about students’ mental images of biological processes. In these interviews, students wondered about what could and could not leave chloroplasts and thylakoids, and what could enter roots. Some students were unsure of what was transported in plants (vs. what is made and used within a cell). This points to the difficulty students have with location and scale.

Students’ Interpretation of Context The context of the question may affect students’ responses. For example, Nehm and Ha (2011) found that students’ explanations of evolutionary trait gains are less sensitive to context, while their explanations for trait loss are more sensitive to context. We explored the effects of context in the cross-over experiment by comparing students’ responses to the corn and maple tree questions about the source of mass gain in growing plants. As seen in Table 4, in both the multiple-choice and essay formats, slightly more students answered the corn question correctly; however, the differences were not statistically significant. There was no difference between the two contexts in the number of students who included both correct and incorrect products on the essay.

We also looked for evidence of context effects in the interviews by comparing interviewees’ responses about the sources of mass gain in radish seeds in a Petri dish and corn. In the radish seed question, eight students eliminated mineral and/or organic substances, because the seeds are in Petri dishes and given only water. Seven students mentioned that corn needs or gets more nutrients (often mentioning fertilizer) than the radish seeds or remarked on the added presence of soil. However, these considerations did not appear to affect their ultimate answer choice. One student stated that corn is not a green plant. He said, “See corn is different because it is obviously a yellowish brown color, so it is not a green plant. He said, “See corn is different because it is obviously a yellowish brown color, so it is not a green plant. He said, “See corn is different because it is obviously a yellowish brown color, so it is not a green plant. However, there were no significant differences in students’ understanding of natural selection found that an open-ended interview reveals that a student who makes a correct choice on one of the multiple-choice cluster questions is an indicator that the student has difficulty performing the practice(s) required by the question. However, the converse is not true. The multiple-true/false questions, essays, and interviews all indicate that a student who makes a correct choice on a multiple-choice question may have: reasoned as expected, remembered a response without understanding, held a mixture of accurate and inaccurate ideas, or simply guessed. This is in line with what others have found. For example, Nehm and Schonfeld’s (2008) study of students understanding of natural selection found that an open-ended response instrument revealed that students who could identify correct responses in a multiple-choice format also held alternative conceptions.

CONCLUSIONS

The Cluster

Most students chose the same response on the exam and in the interview despite an intervening month, indicating test-retest reliability. Those students whose responses differed had explanations for their changes. The exception was Ivy, who switched from incorrect answers on the exam to correct responses in the interview. She was not questioned about this change in the interview.

If we take the interviews to be the best measure of students’ understanding, the interview results demonstrate the construct validity of the diagnostic question cluster. An incorrect choice on one of the multiple-choice cluster questions is an indicator that the student has difficulty performing the practice(s) required by the question. However, the converse is not true. The multiple-true/false questions, essays, and interviews all indicate that a student who makes a correct choice on a multiple-choice question may have: reasoned as expected, remembered a response without understanding, held a mixture of accurate and inaccurate ideas, or simply guessed. This is in line with what others have found. For example, Nehm and Schonfeld’s (2008) study of students understanding of natural selection found that an open-ended response instrument revealed that students who could identify correct responses in a multiple-choice format also held alternative conceptions.

Alternative Formats

Multiple True/False. While the essay questions and interviews revealed students with both accurate and inaccurate ideas, these formats are time-consuming to administer and grade. The results of the cross-over experiment comparing multiple-choice and multiple-true/false formats indicate that the latter can also identify students with heterogeneous understanding. This machine-readable format holds promise as a diagnostic tool with two caveats. Because students need to respond thoughtfully to each statement, they will require somewhat more time to answer questions in multiple-true/false format. Also, if multiple-choice

<table>
<thead>
<tr>
<th>Incorrect</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incorrect: radish seeds in light</td>
<td>30.1%</td>
</tr>
<tr>
<td>Correct: radish seeds in light</td>
<td>11.1%</td>
</tr>
</tbody>
</table>

*Percentages are shown with 95% confidence intervals. n = 720.*
Of basic knowledge of the flow of matter at the subcellular level, questions 5–9 on mass gain and loss in plants assessed students’ understanding of photosynthesis (and respiration) at the organismal level. Again, students’ performances varied from year to year, but at no time did students appear to have a sound understanding of the big picture. We often assume that students come with this high school–level understanding of the relationships among photosynthesis, respiration, and growth and only devote instructional time to elaborating on it without first clarifying it. For example, in Biology (Campbell and Reece, 2005), the chapter on photosynthesis is 19 pages long. The single page that explains the relationship between photosynthesis and respiration is in the preceding chapter, which is devoted to cellular respiration. The section on plant growth comes 500 pages later and refers to cell division, but to no other cellular functions. Thus, it appears that we are layering details on a weak foundation.

Student performance on the three questions about energy sources for plants is also alarmingly weak. Fewer than one-half of the students answer these questions correctly. This indicates that students’ big-picture understanding of the role of photosynthesis in providing reduced-carbon compounds that can be used as energy sources is as weak as their comprehension of the use of reduced-carbon compounds in plant growth. Again, we are giving students molecular details about a process whose basic functions are unclear to them.

The interviews indicate that many students are trying to memorize rather than understand the information (see responses of Wendy, Olivia, Wren, and Sherman in Tables 5 and 6). Students also draw on personal experience and knowledge. For example, many students mentioned that farmers fertilize their corn crops and then tried to fit this piece of information into their answers. Only Burt appeared to reason systematically and with principles about his answers by indicating that if $C_6H_{12}O_6$ is the product of photosynthesis, he must account for a source for each of the three elements.

Implications for Instruction

Taken collectively, these results indicate that many students lack 1) a big-picture understanding of the role of photosynthesis in plant growth and energy use and 2) the knowledge and/or inclination to use basic principles to organize the large amounts of information we want them to learn about these topics. This means that we are layering molecular details onto a weak or faulty foundation without giving students tools to shore up their understanding. We are guilty of ignoring what is known about how people learn. Redish (1994) summarizes this in a set of principles. “Principle 1: People tend to organize their experiences and observations into patterns or mental models” (Redish, 1994, p. 798). “Principle 2: It is reasonably easy to learn something that matches or extends an existing mental model” (Redish, 1994, p. 801). “Corollary 2.1: It’s hard to learn something we don’t almost already know” (Redish, 1994, p. 801). One implication of this is that “new information should always be presented in a context that is familiar to the reader and that the context should be established first” (Redish, 1994, p. 801). More specifically, the implication for teaching photosynthesis is that instruction should start at the organismal scale, clarifying the role of photosynthesis in plant growth and energy use, and only then move to the cellular and subcellular levels. Put bluntly, this means that in a standard semester-long introductory cell and molecular...
biology course, there may not be time to look at all of the individual reactions of photosynthesis. Even if students could memorize all of the reactions, this knowledge would only paper over, rather than correct, their lack of understanding of the role of photosynthesis.

The data presented here indicate that many students cannot learn all of the individual reactions of photosynthesis. They are struggling with just the inputs and outputs of the light-dependent reactions and the Calvin cycle (see results for questions 1–4 in Supplemental Material A). Since memorization appears to be a common student approach to learning, this probably means that the volume of information is too big to manage. People can remember only a limited number of unrelated pieces of information (Miller, 1956; Redish, 1994; National Research Council, 2000). This implies that students do not have the ability to organize and collate the information by connecting the details to the underlying principles. These results support the argument that the standard photosynthesis content should be organized around a few themes such as matter, energy, and scale. In the language of Redish (1994, p. 802), these themes are equivalent to the “touchstone problems” of physics, problems “students will come back to over and over again in later training. Touchstone problems become the analogs on which they will build the more sophisticated elements of their mental models.”

Common Misconceptions and Principled Reasoning

Common misconceptions about photosynthesis represented by distracters in cluster questions can be categorized according to problems with particular practices associated with principled reasoning.

Misconceptions Connected to Not Tracing Matter

- Gases, such as the CO₂ used in photosynthesis, have little or no mass, are unimportant, or cannot account for the mass gain of photosynthetic organisms.
- ATP for cellular use is a product of photosynthesis.
- Atoms from CO₂ end up in ATP.
- Minerals taken up by the roots make a significant contribution to the mass of the plant.
- ATP (from any source) is moved throughout a plant.

Misconceptions Connected to Not Tracing Energy

- ATP made during photosynthesis circulates throughout the plant.
- Sunlight is converted into sugar.
- To produce ATP, plants use respiration when in the dark and photosynthesis when in the light.
- ATP (from any source) is moved throughout a plant.

Misconceptions Connected to Not Keeping Track of Scale and Location

- ATP (from any source) is moved throughout a plant.
- That plants grow is a sufficient explanation for mass gain, without referencing the source of the matter, the source of the energy, or the processes of photosynthesis.

Note that some misconceptions are associated with problems with more than one practice. For example, “ATP is moved throughout the plant” is associated with all three practices. To understand that ATP is made and used locally, students need to understand the inputs and outputs of photosynthesis—phosphate and ADP are not inputs and ATP is not an output of the overall process. An energy lens also provides insights into why ATP is not transported through the plant. Relative to ADP and phosphate, ATP has a lot of chemical potential energy. Thus, hydrolysis of ATP releases that energy. In other words, ATP is unstable, too unstable to last while it is transported. Using a location lens, the ATP made in the chloroplast stays in the chloroplast and is not available for cellular work.

The implications of viewing students’ conceptual barriers in this way are that, if we can provide students with opportunities to practice using a handful of practices and to become better principled reasoners, they will come to understand photosynthesis better. This will require not only making the principles and practices explicit during instruction, but also orienting all instruction around these ideas. The interviews show that many students approach learning as memorization. Learning to reason will require a whole new approach on their part. The instructor’s role in this significant paradigm shift is to give students sufficient time and opportunities to practice reasoning. As a starting point for instruction focused on principled reasoning, we present the content associated with photosynthesis organized around the three practices (Supplemental Material B).

The Vision and Change report (AAAS, 2010) paints in broad strokes the consensus of the biology community that teaching of biology content should be organized around five powerful ideas. The work presented here examines undergraduates’ understanding of two of those ideas (pathways and transformation of energy and matter; systems) in the context of photosynthesis. Data from the use of the diagnostic cluster questions in multiple formats indicate that very few students reason about photosynthesis in this principled way. The work presented here also illustrates why traditional course structures that focus on content coverage have not shown any big gains in student achievement. Without an understanding of, and ability to apply, these principles, students have no foundation on which to build more elaborate and detailed understanding.

ACKNOWLEDGMENTS

We thank Ronald Patterson for his careful reading of the manuscript and helpful comments. The authors gratefully acknowledge the contributions of Chris Wilson, Amy Lark, Duncan Sibley, and the rest of Center for Research on College Science Teaching and Learning community. We also thank Jean Beland for assistance in manuscript preparation. This research was supported by grants from the National Science Foundation (DUE-0243126 and DUE-0736947). The views presented here represent those of the authors and not of the funding agencies.

REFERENCES


Miller GA (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol Rev 63, 81–97.


Our study explored the prospects and limitations of using machine-learning software to score introductory biology students’ written explanations of evolutionary change. We investigated three research questions: 1) Do scoring models built using student responses at one university function effectively at another university? 2) How many human-scored student responses are needed to build scoring models suitable for cross-institutional application? 3) What factors limit computer-scoring efficacy, and how can these factors be mitigated? To answer these questions, two biology experts scored a corpus of 2556 short-answer explanations (from biology majors and nonmajors) at two universities for the presence or absence of five key concepts of evolution. Human- and computer-generated scores were compared using kappa agreement statistics. We found that machine-learning software was capable in most cases of accurately evaluating the degree of scientific sophistication in undergraduate majors’ and nonmajors’ written explanations of evolutionary change. In cases in which the software did not perform at the benchmark of “near-perfect” agreement (kappa > 0.80), we located the causes of poor performance and identified a series of strategies for their mitigation.

Machine-learning software holds promise as an assessment tool for use in undergraduate biology education, but like most assessment tools, it is also characterized by limitations.

INTRODUCTION

In large introductory biology classes throughout the United States, multiple-choice (MC) formats typify both formative assessments (e.g., clicker questions, concept inventories) and summative tests (e.g., midterm and final exams; see Wood [2004] and Smith et al. [2008]). While there is little doubt among educators that MC formats in general are capable of providing cost-effective, reliable, and valid inferences about student knowledge and misconceptions in many content areas, not all types of student learning outcomes may be measured using MC formats (reviewed in American Association for the Advancement of Science [AAAS, 2011] and Nehm et al. [in press]). Moreover, despite generating useful assessment information, MC tests may also produce unintended, and rarely considered, negative consequences for learners, such as the generation of false knowledge (Mandler and Rabinowitz, 1981; Roediger and Marsh, 2005; Butler et al., 2006; Kang et al., 2007). Additionally, many MC tests are most conducive to detecting novice or expert (incorrect or correct) models of student thinking, whereas a large body of work in cognitive science indicates that many students construct mixed models of naïve and informed scientific information as they learn (e.g., Vosniadou [2008]; Opfer et al. [2011]); right or wrong options—the staple of MC tests—may limit the valid measurement of student learning gains (Nehm and Schonfeld, 2008; Nehm and Ha, 2011; Neumann et al., 2011).

Collectively, these and many other limitations of MC formats should motivate biology educators to 1) develop and deploy a more diverse array of high-quality assessment methods and 2) measure a more expansive range of student...
knowledge, skills, and learning outcomes (e.g., AAAS [2011, p. 17]; Nehm et al. [in press]). The purpose of our study was to investigate the prospects and limitations of implementing new assessment methods in introductory biology—specifically, computerized scoring of short-answer scientific explanations. Can successful application of these innovative methods at one university be generalized, and if not, why not? What are the implications for adopting computerized-scoring programs as assessment tools in introductory biology and biology education research?

BACKGROUND

Open-Response Assessments in Biology Education

Educational researchers emphasize that assessments should be built upon and aligned with what we know about student learning and cognition (National Research Council [NRC], 2001). One major advance in our understanding of student learning is that learners do not progress directly from novice to expert levels; rather, the pathways of knowledge growth in biology (and other domains) are highly inefficient and involve integrating scientific ideas into naïve knowledge frameworks, generating heterogeneous mental models, or building coexisting models (Nehm and Schonfeld, 2008, 2010; Vosniadou, 2008; Kelemen and Rosset, 2009; Evans and Lane, 2011). These so-called mixed or synthetic models may persist for long periods of cognitive development (Vosniadou, 2008) and even through the years of college instruction (Nehm and Reilly, 2007).

If our biology assessments are intended to measure progress in student reasoning and build upon findings from educational research, our assessment items (whether formative or summative, MC or open response) must permit at least three general reasoning categories as options: 1) exclusively naïve answer choices; 2) assemblages of mixed or synthetic answer choices; and 3) exclusively scientific answer choices. Currently, many MC diagnostic and summative biology assessments (and concept inventories) contain option types 1 and 3 (novice and expert, respectively), despite mounting evidence in some biology domains that most students harbor “mixed models” of biological concepts (Nehm and Haertig, in press). Moreover, differently trained graders often disagree about the scores that should be given to a response, requiring additional training time to equalize scoring among raters. Reliable and consistent scoring of constructed-response items cannot be solved by having one human grader score all of the responses; grading fatigue and changes in grading precision are well-known limitations of human scoring (Nehm and Haertig, in press). A more serious issue with human scoring of written responses is the persistent problem of grading subjectivity and consequent reliability threats; such problems are often introduced by the need for many different human graders to score large data sets, such as those generated in undergraduate biology courses (Nehm and Haertig, in press). Moreover, differently trained graders often disagree about the scores that should be given to a response, requiring additional training time to equalize scoring among raters. Reliable and consistent scoring of constructed-response items cannot be solved by having one human grader score all of the responses; grading fatigue and changes in grading precision are well-known limitations of human scoring (Nehm and Haertig, in press). Thus, many long-standing problems have limited the use of open-response formats.

One final argument for the inclusion of open-response assessments in introductory biology is that they better align with most real-world experiences than do MC formats. Increasingly, college graduates are expected to perform non-routine tasks that cannot be automated, digitized, or outsourced (Nehm et al., in press). From an educational point of view, deploying assessment tasks that model authentic problem-solving environments would help reinforce for students which types of performances are most highly valued by biology educators, and which types of evaluations they are likely to experience postgraduation (e.g., production vs. selection tasks). Overall, while MC assessments should remain in biology educators’ assessment toolboxes, the many advantages of open-response formats call for their greater inclusion. Practical limitations have prevented the wider use of open-response assessments, but recent technological advances are beginning to change this situation.

Computer-Assisted Scoring Tools

The increasing use of computer-assisted scoring (CAS) in many educational contexts has been motivated by the numerous constraints that characterize human scoring of constructed-response (e.g., short-answer, essay) items. Some of the most obvious limitations are the large amounts of time, money, and expertise needed to score such responses, and the consequent delayed feedback to test takers (Nehm et al., in press). A more serious issue with human scoring of written responses is the persistent problem of grading subjectivity and consequent reliability threats; such problems are often introduced by the need for many different human graders to score large data sets, such as those generated in undergraduate biology courses (Nehm and Haertig, in press). Moreover, differently trained graders often disagree about the scores that should be given to a response, requiring additional training time to equalize scoring among raters. Reliable and consistent scoring of constructed-response items cannot be solved by having one human grader score all of the responses; grading fatigue and changes in grading precision are well-known limitations of human scoring (Nehm and Haertig, in press). Thus, many long-standing problems have limited the use of open-response formats.

Fortunately, the rapid pace of developments in computer technology and text analysis software has made CAS tools more economical and accessible to educators. Consequently, many of the aforementioned limitations of human scoring have been investigated empirically using a variety of different software tools. This work has demonstrated that computer software can be “trained” to score constructed-response items as accurately and reliably as human raters (Page, 1966; Yang et al., 2002; Shermis and Burstein, 2003). Indeed, the Educational Testing Service and many other large companies now employ CAS methods in large-scale, high-stakes, standardized exams (Powers et al., 2002). Examples of these CAS tools include C-rater (Sukkarieh and Bolge, 2008), E-rater (Burstein, 2003; Williamson, 2009), and Intelligent Essay [AAAS, 1994]; the National Science Education Standards [NRC, 1996]; Taking Science to School [Duschl et al., 2007]; and Vision and Change in Undergraduate Biology Education [AAAS, 2011]; Braaten and Windschitl (2011)]. The ability to generate scientific explanations can only be assessed using open-response formats.
Carnegie Mellon University (Mayfield and Rosé, 2010a, b). SIDE is a freely available software package distributed by software package sold by a private company (IBM), whereas SIDE is a freely available software package distributed by Carnegie Mellon University (Mayfield and Rosé, 2010a, b). Although the performance efficacy of both computer programs has been demonstrated using samples of undergraduate science students (Ha and Nehm, 2011), the two programs differ in the ways in which they approach CAS, as well as in the methods used to perform scoring (for a review of these differences, see Haudek et al. [2011] and Nehm et al. [in press]). SIDE is capable of creating scoring algorithms automatically when provided with a sufficiently large set of human-scored data for training and validation of scoring algorithms. Because we had a large sample of human-scored, short-answer responses for this study, our work on CAS was ideally suited to using SIDE.

SIDE combines a natural language processing (NLP) engine for parsing text, along with a set of machine-learning algorithms for classifying text (see Witten and Frank [2005] for more details). Analyzing text responses using SIDE has two main steps: 1) defining the filters necessary for capturing the structure of the text and 2) specifying the summaries to be displayed and extracting the needed subsets (for details, see Mayfield and Rosé [2010a, b] and the supplemental material). Operationally, SIDE uses corpora of students’ constructed responses that have been scored by humans to detect text patterns associated with the presence or absence of particular scientific concepts as measured by expert raters. For instance, terms such as “mutation,” “genetic change,” or “change in DNA” are indicative of the presence of variation, which is one of the key concepts necessary for explaining evolutionary change (Nehm and Reilly, 2007). Student responses, and the associated expert scores, are used as input to SIDE. Since we focused on five key concepts in this study, we needed to input a set of human-scoring information on whether or not the student text included a particular concept or not for each of the five key concepts of evolution that we investigated (see Human Scoring of Explanations of Evolutionary Change).

SIDE provides a number of interactive tools for refining and improving the accuracy of the predictions by allowing the user to examine cases where the machine-learning model misclassified (either incorrectly classifying a response as containing the concept when it did not, or failing to classify a response as containing a concept when it in fact does). SIDE can save the scoring model and apply it to new text to predict human scoring (Mayfield and Rosé, 2010a, b).

In this study, we used initial data sets, which had been scored by two biology experts, to train SIDE. We then used new, expert-scored data to validate the accuracy of the SIDE-scoring models. If this cross-validation approach is successful, this provides evidence that we can predict human scoring of new sets of student responses to the same questions with as much confidence as we would have using human raters.

**RESEARCH QUESTIONS**

Our study explores three research questions:

1. Are scoring models built using machine learning generalizable across colleges and courses (majors and nonmajors at different universities)? In other words, do scoring models built using student responses at one school function effectively at other schools?
2. How many human-scored student responses are needed to effectively build scoring models? To what degree does sample size impact computer-scoring success?
3. What factors limit computer-scoring efficacy, and how can these factors be mitigated to enable scoring models to be used in introductory biology courses across universities?

**METHODS**

**Sample**

To answer our research questions, we utilized three samples of undergraduate students enrolled in biology coursework (Table 1): 1) nonmajors enrolled in introductory biology at Ohio State University (OSU; 264 students/1056 written explanations); 2) nonmajors enrolled in introductory biology at Michigan State University (MSU; 146 students/584 written explanations); and 3) biology majors enrolled in introductory biology at MSU (440 students/1760 written explanations). Student responses were gathered using two online survey systems (ACS and LONCAPA; for details, see www.evolutionassessment.org and www.lon-capa.org, respectively).

We only included responses from individuals who completed four survey items each with responses of more than five words. The number of participants in the OSU nonmajor sample who completed all four items (see following section) was 358 (77.7% of total participants). Given the significant labor involved in scoring open-response items, we randomly selected a subset of 264 students (1056 responses) from this sample. We sampled the MSU data using the same approach. The number of participants in the MSU nonmajors sample who completed all four items with more than five words was
146 (90.7% of total participants). Finally, the number of participants in the MSU biology major sample who completed all four items with more than five words was 440 (90.0% of total participants). Because of the time, money, and expertise required to score student responses, we randomly selected a subset of responses \( (n = 500) \) from each sample (OSU nonmajors, MSU nonmajors, and MSU majors; 1500 responses total) for the first research question. For our second research question, we scored more than twice as many responses from the OSU corpus \( (n = 1056) \).

**Items Used to Generate Explanations of Evolutionary Change**

Our response corpus was composed of student explanations from four open-response items about evolutionary change. This assessment format has been employed in biology education research for more than 25 yr (e.g., Clough and Driver [1986]) and has been shown to generate reliable and valid inferences about students’ reasoning regarding evolutionary change (Bishop and Anderson, 1990; Nehm and Schonfeld, 2008; Nehm and Ha, 2011). Instrument items were isomorphic (’’How would biologists explain how a living X species evolved from an ancestral Y?’) but differed in specific taxa and traits (i.e., X and Y).

Assessment item features, such as trait functionality and organism type, are known to influence students’ reasoning regarding evolutionary change (Nehm and Ha, 2011; Opfer et al., 2011). Consequently, we standardized our prompts to include only animal examples and functional traits (e.g., fins, wings). Moreover, because the familiarity of taxa and traits is also known to influence students’ reasoning regarding evolutionary change (Opfer et al., 2011), item taxa and traits were constrained by their overall familiarity. To do so, we used the frequencies of “organism + trait” in Google rankings as a proxy for familiarity (Nehm, Beggrow, et al., in press). Specifically, taxon/trait combinations included: shrew incisors, snail feet, fish fins, and fly wings.

Although students’ short-answer explanations of evolutionary change varied in length, the average number of words did not differ among items (analysis of variance [ANOVA], \( F = 3.04, P > 0.01 \)). Specific item lengths were 1) shrew: mean = 45.5, SD = 30.3, minimum = 6, maximum = 430; 2) snail: mean = 42.9, SD = 27.9, minimum = 6, maximum = 429; 3) fish: mean = 42.5, SD = 26.3, minimum = 6, maximum = 202; 4) fly: mean = 41.6, SD = 26.5, minimum = 6, maximum = 209.

**Human Scoring of Explanations of Evolutionary Change**

Undergraduate students are known to recruit a diverse array of cognitive resources to build explanations of evolutionary change and solve evolutionary problems (Nehm, 2010). These resources may include, for example, well-structured scientific schemas, as natural selection; fragmented mental models built using mixtures of scientific and naïve knowledge elements; or naïve explanatory models (Nehm and Ha, 2011). Given such diversity, it is most practical for assessment purposes to capture the existence of constituent explanatory elements in students’ explanatory models (cf. Nehm and Haertig [in press]). For our study of automated computer scoring, two trained human raters scored all student responses for five key concepts of natural selection that were outlined by Nehm and Reilly (2007), described by Nehm and Schonfeld (2008), and codified in the scoring rubrics by Nehm et al. (2010a). It is important to emphasize that these concepts are central to the construct of natural selection, and necessary for explaining the operation of natural selection (Nehm and Schonfeld, 2010). Thus, the elements selected for scoring are not trivial or superficial aspects of reasoning regarding evolutionary change, and are associated with explanatory competence, as measured by clinical oral interviews (Nehm and Schonfeld, 2008). These key concepts included: 1) the presence and causes of variation (mutation, recombination, sex), 2) the heritability of variation, 3) competition, 4) limited resources, and 5) differential survival. It is important to note that scoring of short-answer explanations was dominated by the recognition of collections of key terms and short phrases, rather than elaborate grammatical expressions (see Nehm et al. [2010a] for details). Nevertheless, scoring was performed such that only accurate expressions counted for the “presence” of a concept; students’ faulty expressions about heredity, for example, would not count as the presence of the key concept of heredity.

A series of studies has demonstrated that the coding rubrics used to score the short-answer explanations of evolutionary change are sufficiently clear to produce high levels of human interrater scoring agreement with moderate training (Nehm and Reilly, 2007; Nehm and Schonfeld, 2008; Nehm et al., 2009a; Nehm et al. 2010b; Nehm and Ha, 2011; Nehm and Haertig, in press; Nehm et al., in press). In past studies, kappa agreement coefficients between human scorers with limited training ranged from 0.69 to 0.95, with an average of 0.86 (e.g., Nehm and Haertig, in press). In the present study, two biology experts (who have scored several thousand explanations of evolutionary change and have used the rubrics of

<table>
<thead>
<tr>
<th>Institution</th>
<th>Major</th>
<th>Participants (n)</th>
<th>Ethnicity (%)</th>
<th>Gender (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>White</td>
<td>Minority</td>
</tr>
<tr>
<td>OSU</td>
<td>Nonmajor</td>
<td>264</td>
<td>79.1</td>
<td>14.4</td>
</tr>
<tr>
<td>MSU</td>
<td>Nonmajor</td>
<td>146</td>
<td>66.4</td>
<td>13.7</td>
</tr>
<tr>
<td>MSU</td>
<td>Major</td>
<td>440</td>
<td>79.1</td>
<td>11.8</td>
</tr>
</tbody>
</table>

\( ^a \)OSU = Ohio State University; MSU = Michigan State University.

\( ^b \)Note that \( n \) refers to subsampled data sets (see Sample).
Nehm et al. [2010a] for more than 2 yr) evaluated all student responses with very high agreement levels. The kappa reliability coefficients ($\kappa = 1056$) for the present study were: 0.995 for variability, 1.000 for heritability, 1.000 for competition, 1.000 for limited resources, and 0.988 for differential survival. In the rare cases of disagreements between the two human raters, consensus scores were reached via deliberation. These final consensus scores were used in all subsequent analyses of human–computer scoring correspondence.

**Human–Computer Correspondence Measures**

We used Cohen’s kappa to quantify the magnitude of human–computer scoring correspondence (Bejar, 1991). Cohen’s kappa values range from 0.0 to 1.0, and are commonly used to quantify levels of agreement among human raters, or between human and computer rating scores (Landis and Koch, 1977; Nehm and Haertig, in press). Landis and Koch (1977) introduced three general agreement levels for kappa statistics that we follow in our study: values between 0.41 and 0.60 are considered “moderate”; values between 0.61 and 0.80 are considered “substantial”; and those between 0.81 and 1.00 are considered “near perfect.” We also report specific kappa values for all analyses, given the subjective nature of these categorical distinctions.

**ANALYSES**

Our first research question explores SIDE performance at detecting individual key concepts of natural selection in students’ written responses. For each of our analyses, and for each key concept, two categories of agreement statistics are reported: 1) scoring-model training agreement values and 2) scoring-model cross-validation agreement values. The training agreement values are generated when SIDE first attempts to construct a scoring model using the corpus of human-scored student answers; that is, SIDE attempts to “learn” from the human-scoring patterns and builds a computational model that can account for these patterns. Then, SIDE examines the efficacy of this scoring model by calculating how well the model scores the same responses from which it learned. Kappa and percentage agreement values (which are automatically generated by the SIDE program) enable researchers to judge the strength of the machine-learning model and consider whether it is worthy of use on a new data set. In situations where the training kappa values are “substantial” (> 0.60), the SIDE-generated scoring model is then applied to a new corpus of human-scored responses to determine whether the scoring model functions effectively with a new response corpus (that is, the training model is tested). Even if a training model performs admirably, this does not necessarily mean that it will be effective at scoring a new response corpus; both training and cross-validation performances need to be evaluated. Model cross-validation efficacy (also measured using kappa and percentage agreement statistics) must be performed manually (in our case, using SPSS, version 19.0). Cross-validation kappa values and percentage agreement values enable us to determine whether the SIDE-generated scoring models are likely to effectively score additional student responses.

In addition to exploring training and cross-validation performance of SIDE for each individual key concept, we also explored composite measures of students’ explanations of evolutionary change. Key concept score (KCS) is a composite measure of the number of scientific concepts employed in an explanatory context (Nehm and Reilly, 2007). Given that KCS has been used in prior research on learning gains (Nehm and Reilly, 2007), we examined how well SIDE performed relative to human expert scorers for KCS. Specifically, we used Pearson correlation statistics (in SPSS, version 19.0) to test for significant associations between human and computer scores of both variables (in contrast to the single-concept agreement statistics discussed above).

Our second research question examined to what degree sample size influences SIDE scoring–model performance. Specifically, we trained SIDE on two different corpora: 1) 500 responses from OSU students and 2) 1056 responses from OSU students. We then tested the two different scoring models on 1) the MSU nonmajor response corpus and 2) the MSU biology major response corpus. We calculated kappa and percentage agreement statistics to evaluate the influence of the training-sample size on SIDE scoring–model performance. All statistical tests were performed in SPSS, version 19.0.

Our third research question explored the factors that limit SIDE scoring–model performance; that is, why, in some cases, do scoring models fail to function at the desired near-perfect (kappa values > 0.80) agreement levels? Are such disagreements the products of the students (majors, nonmajors) and how they explain evolutionary change; the universities (OSU, MSU); the sample sizes; the scoring models; or combinations of these factors? This research question required examining all of the instances in which SIDE-generated scores and human-generated scores did not match, and attempting to identify the factors that contributed to score mismatches. After locating the likely source of scoring disagreements, we explored whether there were ways to mitigate these performance limitations so that future work would be more effective.

**RESULTS**

**Students’ Explanations of Evolutionary Change**

To provide readers with a sense of the types of explanations of evolutionary change that undergraduate students’ generate, four unedited student responses were extracted from the response corpus (see Table 2). As is apparent, student explanations of evolutionary change vary in length (for details, see Items Used to Generate Explanations of Evolutionary Change), sophistication, scientific accuracy, and scientific complexity. Adjacent to the responses in Table 2 are two columns indicating the numbers and types of key concepts detected in each response (see scoring methods, in Methods, for details). Note that in the present study we investigated only the magnitudes of accurately expressed scientific concepts in student responses, not naïve ideas or misconceptions. Computer scoring of other explanatory elements is the focus of ongoing research.

**Testing the Impact of Training Corpus on Scoring Success**

Our first analyses explored whether training SIDE using different human-scored corpora had an impact upon
scoring-model success (Figure 1). Six tests were performed (Figure 1, A–F) for each key concept of evolution (e.g., variation, heredity, etc.). For the majority of these tests, the scoring agreements reached or exceeded near-perfect kappa values (18/30 tests) and percentage agreements above 90% (24/30 tests). Three key concepts—variation, heredity, and limited resources—were detected at near-perfect levels regardless of the training or cross-validation samples used. In contrast, competition and differential survival were very sensitive to training and cross-validation samples; in only two of the 12 tests did they reach near-perfect kappa agreement levels (Figure 1, left). While raw percentage agreement values were robust for four of the five concepts (the exception being differential survival), these values do not take into account chance agreements. The dramatic difference between these two agreement statistics for competition suggests that sample size is contributing to these patterns, as we discuss in Training Sample Sizes and Scoring Success. Overall, the most significant factor influencing the performance of the SIDE-generated scoring models was not the training or cross-validation corpus per se, but rather, concept-specific factors.

**Concept Frequencies in Different Samples**

As shown in Figure 2, human-identified frequencies of key concepts (blue bars) are in close alignment with computer-identified frequencies (red and green bars) for all samples (the different rows: OSU nonmajors, MSU majors, and MSU nonmajors). In addition, different SIDE training sets (i.e., MSU majors, MSU nonmajors) did not generate substantially different scoring frequencies of key concepts in comparison with human-generated scores (compare the different colored bars for each concept within each row). Small differences are apparent, however, among scores for differential survival in the OSU nonmajor sample (top row, right) and variation in the MSU major and nonmajor sample (middle and bottom rows, left).

One of the most striking patterns across all samples is that introductory biology students rarely used the concept of competition in their explanations of evolutionary change. In contrast, differential survival was used by a majority of students in all samples. Differences in the frequencies of key concept use were also apparent between samples (compare the rows in Figure 2): almost twice as many responses from the MSU samples employed the concept of variation, compared with the responses in the OSU samples (bottom two rows vs. top row, left). In addition, MSU majors used the concept of limited resources more often than the other groups. Overall, Figure 2 demonstrates that differences in the frequencies of particular concepts vary across samples and schools, but the SIDE-generated scoring model was able to detect these differences. The extremely rare occurrence of competition (Figure 2) was associated with poor model performance for this concept (Figure 1).

**Training Sample Sizes and Scoring Success**

Given that some key concepts were less common in the training and cross-validation data sets than other concepts (e.g., competition vs. variation), we investigated the impact of sample size on scoring-model efficacy. For each key concept (e.g., variation, differential survival) we performed two experiments. In the first, we trained SIDE using 500 human-scored responses from a sample of OSU nonmajors; and in the second, we trained SIDE using a sample more than twice as large: 1056 human-scored responses from OSU nonmajors.

### Table 2. Selected examples of students’ written explanations of evolutionary change and corresponding human and computer scores

<table>
<thead>
<tr>
<th>Taxon/trait/polarity</th>
<th>Student’s explanation of evolutionary change</th>
<th>Human score (number of key concepts)</th>
<th>Computer score (number of key concepts)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shrew incisors</td>
<td>“Incisors may have developed on shrews due to a genetic mutation [Variation]. An offspring of a normal shrew may have had a mutated baby that had incisors, or some earlier form of incisors. The incisors would have given the new shrew an advantage in acquiring food [Limited resources] and reproducing, so it would have a higher fitness [Differential survival] leading the incisor trait to be passed on to other generations [Heredity]. As the trait will then develop with each generation due to variation involving the trait and the levels of success attached to each variant.”</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Snail feet</td>
<td>“They would explain that once all the snails had small feet. Then one day there was a mutation [Variation] that produced a snail with a large foot. The snail with a large foot was better able to produce more offspring [Differential survival] in the environment passing on his trait [Heredity].”</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Fish fins</td>
<td>“There was a random change in the DNA sequence [Variation] of the fish that coded for the production of the fin. Nonrandom mating could have occurred with females selecting males with fins as partners, which disrupts HW equilibrium and leads to the evolution of the fin because the fish with fins are better able to produce viable offspring [Differential survival].”</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Fly wings</td>
<td>“The evolution of a fly species with a large wing from an ancestral fly with small wings could be through the process of natural selection or from a random mutation [Variation].”</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Figure 1. Magnitudes of agreement among humanscored and computer-scored explanations of evolutionary change from three samples (OSU, MSU non-major, and MSU major). For each of the three samples: n = 500 responses. Five key concepts of evolutionary change were examined separately (e.g., variation, heredity). Arrows indicate which sample was used to train the models and which sample was used to test the models. Kappa values compensate for chance agreements, whereas agreement values are raw percentages. (A) OSU sample model training and MSU sample nonmajor model cross-validation; (B) MSU nonmajor sample model training and OSU sample model cross-validation. (C) OSU sample model training and MSU nonmajor model cross-validation. (D) MSU major sample model training and OSU sample model cross-validation. (E) MSU major sample model training and MSU nonmajor model cross-validation. (F) MSU nonmajor model training and MSU major sample cross-validation.

Figure 2. Frequencies (0–100%) of key concepts among samples and between human- and computer-generated scores. Blue bars = human-detected frequencies; red bars = frequencies detected using the MSU major computer-generated scoring model; and green bars = the frequencies detected using the MSU nonmajor computer-generated scoring model. In each of the three samples (OSU nonmajor; MSU major; MSU nonmajor), 500 responses were used. Error bars represent the SEM.
The two resulting SIDE-generated scoring models were applied separately to: 1) a corpus of MSU nonmajors’ written explanations and 2) a corpus of MSU biology majors’ written explanations (see Methods). As above, kappa agreement statistics and raw agreement percentages were calculated for all comparisons. Figure 3 illustrates the results of both experiments.

The scoring models built from the larger corpus produced higher correspondences with expert human raters in nine out of 10 tests; the exception was competition in the MSU biology major sample (Figure 3). Given that the smaller training corpus (n = 500) produced near-perfect correspondence with human raters in most tests, doubling the training size bumped only one concept—competition in the MSU nonmajor sample—to the desired benchmark (kappa > 0.80). Although the larger corpus did improve model performance for differential survival (Figure 3), in many cases it did not meet our benchmark using either training corpus (n = 500 or 1056). In terms of raw percentage agreements, the larger corpus did not always produce improved results; in fact, the smaller corpus in many cases produced slightly higher agreement percentages. Nevertheless, in tests of model performance, 17 out of 20 comparisons of computer and human agreement reached or exceeded 90%. Additionally, using the large training data set, 9/10 analyses produced results that were detected at or above the kappa benchmark of 0.80.

Overall, in nearly all cases, doubling the training corpus improved model performance, but not substantially. The most dramatic improvement was seen in the detection of competition in the MSU nonmajor sample. Thus, the frequencies of particular concepts in the training corpus must be considered, not just overall sample size (see Figure 2 “Human scoring”).

**Explanatory Structures**

In addition to comparing individual key concept detection between human- and computer-scored explanations, it is useful to examine how students collectively assemble these concepts into explanatory structures (Figure 4). One approach for representing these explanatory interrelationships is to use concept-association diagrams (Nehm and Ha, 2011). Figure 4 illustrates both the frequency (the size of the circles) and co-occurrences of concepts (the thickness of the gray lines) in students’ explanations of evolutionary change. For instance, approximately 20% of OSU nonmajors used both concepts of variation and differential survival in their responses (see connecting lines in Figure 4). Each row in Figure 4 compares explanatory structures between human expert raters (left) and SIDE-scoring patterns (right) for a particular student sample (e.g., top: OSU nonmajors; bottom: MSU biology majors). As is apparent from the figure, results for students’ knowledge networks are remarkably similar, regardless of whether they were scored by humans or computers.

Figure 3. Cross-validation of the impact of training sample size on model performance. Four samples were used in the analysis (OSU nonmajors: n = 500; OSU nonmajors: n = 1056; MSU nonmajors: n = 500; and MSU majors: n = 500). Five key concepts of evolutionary change were examined separately (e.g., variation, heredity). Arrows indicate which sample was used to train the models and which sample was used to test the models. Kappa values compensate for chance agreements, whereas agreement values are raw percentages. (A) OSU sample (n = 500) training and MSU sample nonmajor cross-validation. (B) OSU sample (n = 1056) training and MSU sample nonmajor cross-validation. (C) OSU sample (n = 500) training and MSU major cross-validation. (D) OSU sample (n = 1056) training and MSU sample major cross-validation.
Comparing the columns in Figure 4 also reveals that SIDE-generated scoring models can detect different explanatory structures across student samples, and these patterns closely mirror the findings reported in Figure 2. The human-generated scores, for example, demonstrate that MSU majors and nonmajors used the concept of variation much more frequently than OSU nonmajors. SIDE produced the same patterns. Interestingly, human scorers also determined that MSU nonmajors used the concepts of variation and heredity more frequently than MSU majors; SIDE detected these patterns. MSU biology majors used the concept of limited resources much more frequently than MSU nonmajors, and this is also indicated in the SIDE-generated scores. While differences in the explanatory structures among majors, nonmajors, and institutions are interesting, the important point we wish to emphasize in Figure 4 is that SIDE-generated scores, which took minutes to generate, are in remarkable alignment with the patterns generated by humans during weeks of painstaking grading.

In addition to examining patterns of correspondence between human- and computer-generated visual representations of explanatory structure, Nehm and Reilly (2007) used KCS to quantify the number of different scientifically accurate evolutionary concepts that students use to explain evolutionary change in a prompt. Table 3 illustrates statistically significant ($P < 0.001$) and robust ($r = 0.79$ to 0.87) associations between human- and computer-generated KCS for all comparisons. Thus, using approaches for measuring student knowledge of evolution previously established in the literature (Nehm and Reilly, 2007; Nehm and Ha, 2011), SIDE-generated scoring models produce patterns equivalent to those derived from human raters.
In all cases, associations were strong and significant (*P < 0.001). KCS represents the number of different scientific concepts in a prompt. For details, see Methods and Nehm and Reilly (2007).

Table 3. Correlation coefficients between human-scored and SIDE-scored student explanations for KCS

<table>
<thead>
<tr>
<th>Training sample</th>
<th>Testing sample</th>
<th>Human vs. SIDE KCD correlation (*P &lt; 0.001)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSU nonmajor</td>
<td>MSU nonmajor</td>
<td>0.79**</td>
</tr>
<tr>
<td>(n = 500)</td>
<td>(n = 500)</td>
<td></td>
</tr>
<tr>
<td>MSU nonmajor</td>
<td>OSU nonmajor</td>
<td>0.80**</td>
</tr>
<tr>
<td>(n = 500)</td>
<td>(n = 500)</td>
<td></td>
</tr>
<tr>
<td>OSU nonmajor</td>
<td>MSU major</td>
<td>0.87**</td>
</tr>
<tr>
<td>(n = 500)</td>
<td>(n = 500)</td>
<td></td>
</tr>
<tr>
<td>MSU major</td>
<td>OSU nonmajor</td>
<td>0.85**</td>
</tr>
<tr>
<td>(n = 500)</td>
<td>(n = 500)</td>
<td></td>
</tr>
<tr>
<td>MSU major</td>
<td>MSU nonmajor</td>
<td>0.82**</td>
</tr>
<tr>
<td>(n = 500)</td>
<td>(n = 500)</td>
<td></td>
</tr>
<tr>
<td>MSU nonmajor</td>
<td>MSU major</td>
<td>0.82**</td>
</tr>
<tr>
<td>(n = 500)</td>
<td>(n = 500)</td>
<td></td>
</tr>
</tbody>
</table>

*In all cases, associations were strong and significant (*P < 0.001). KCS represents the number of different scientific concepts in a prompt. For details, see Methods and Nehm and Reilly (2007).

Table 4. Examples of the types of disagreements between human-scored and computer-scored explanations

<table>
<thead>
<tr>
<th>Scoring pattern</th>
<th>Category</th>
<th>Examples 1 to 5</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive computer score but negative human score</td>
<td>Many key terms used, but important aspects were missing</td>
<td>(1) “The original Shrew, who didn’t have incisors, may have not been a fit species. Meaning, they may have not been reproducing enough to pass on their traits. Another reason would be a mutation that survived. What I mean is that a few shrews may have developed an allele mutation that resulted in the shrews developing incisors. After the mutation, it was passed on and probably survived because of natural selection, sexual selection, or artificial selection [more fit, survive and reproduce more].”</td>
<td>Human augmentation of SIDE-scoring models</td>
</tr>
<tr>
<td>Negative computer score but positive human score</td>
<td>Very uncommonly used expression</td>
<td>(3) “The fish was filling a niche in an area that required a fish with smaller fins. Generations passed and a mutant gene for a fish with smaller fins did well and its offspring did well through time a new species was born.”</td>
<td>Increase concept frequencies in training sample; human augmentation of SIDE-scoring models</td>
</tr>
<tr>
<td>Complex expressions</td>
<td></td>
<td>(4) “Variation of living fish species may leads [sic] to random mutation. It creates new sequences of DNA that will code for new or different protein. This protein helps [sic] in the creation of a new living fish species with wings in this situation. This species then reproduce [sic] and evolve through some time. It may help to explain how a new living fly species with wings evolve [sic] even though it is originally from an ancestral fly species that lacked wings.”</td>
<td>Human augmentation of SIDE-scoring models</td>
</tr>
<tr>
<td>Spelling errors and spacing errors</td>
<td></td>
<td>(5) “Preditor [predator], survive [survive], springoffs [offspring], foodso [food so]”</td>
<td>Incorporate a spell-check program during data collection</td>
</tr>
</tbody>
</table>

*Categories: types of scoring problems; examples: specific student responses; solutions: approaches used to correct the computer–human disagreement.

Factors Limiting Computer-Scoring Success

Although SIDE and its machine-learning algorithms were shown to be highly effective at scoring the accuracy and complexity of undergraduates’ explanations of evolutionary change, our studies revealed several limitations, which are summarized in Table 4. The key factors that limited the efficacy of computer scoring included: misspellings; nonadjacent key terms; very uncommon concept frequencies; and the diversity of expressions that students used to represent particular concepts.

Spelling and spacing errors produce human–computer score disagreements. While our human raters easily understood what students were attempting to explain when they wrote “preditor” [predator], the computer was not able to do so. Misspelled words such as “servive” [survive], “springoffs” [offspring], and “foodso” [food so] also produced misclassifications in our study (see Table 4).

We also found that when student responses included key terms suggestive of a concept, but the words constructing the concept were scattered throughout the written response...
(conveying a very different meaning), the SIDE-scoring models often mistakenly linked these words together and considered the concept to be present. For example, in Table 4, example 2, the scoring model identified text elements characteristic of the concept of differential survival (“survive longer”) to be present, even though the response included the words “survive” and “longer” in separate sentences.

Complex expressions, or very long sentences, also posed problems for the software. For example, one student (correctly) explained differential survival in the following way: “the fish with large fins is not suitable for the living environment in that specific area, so the fish with smaller fins survive and reproduce.” Because expressions like this were rare in the student response corpora, and did not contain explicit language about differential survival, the scoring models failed to detect them. Fortunately, expressions like this were uncommon in the samples.

Low concept frequencies also prevented the machine-learning algorithm from building successful scoring models; without enough positive cases to analyze, the computer failed to annotate new cases appropriately. Competition serves as an example of a concept that was very rarely used by students to explain evolutionary change; only about 10 instances of competition were found among 500 written explanations. In addition to low concept frequencies, unusual expressions also lead to misclassifications. The term “mutant gene,” while clearly identifiable as a “cause of variation” by a biologist, was too rare to be incorporated into the scoring models (Table 4).

We found that the concept of differential survival was influenced by the frequencies with which particular samples used particular terms. While the kappa values between scoring models built using the OSU nonmajor and the MSU major sample were nearly 0.80 (near-perfect), the kappa values between scoring models built by the OSU nonmajor and MSU nonmajor samples and between the MSU major and MSU nonmajor samples did not meet this benchmark (0.60). The cause of this pattern appears to be that the language patterns that MSU nonmajors used were somewhat different from the language patterns that the OSU nonmajor and MSU major samples used to describe differential survival. For example, the term “differential” (such as “differential reproduction rate” or “differential survival success”) was observed 12 times among 500 responses in MSU nonmajor sample, whereas the term “differential” was observed only three times among 500 responses in the OSU nonmajor sample and only once among 500 responses in the MSU major sample. Consequently, the program incorporated “differential” as a diagnostic term for the MSU nonmajor scoring model, but not for the other samples.

DISCUSSION

While CAS is becoming increasingly common throughout the educational hierarchy (Nehm et al., in press), biologists have been slow to make use of this technological innovation. Two recent studies by Nehm and Haertig (in press) and Nehm et al. (in press) tested the efficacy, respectively, of SPSSTA, version 3.0 (Galt, 2008) and SIDE (Mayfield and Rosé, 2010a,b). Using large samples of undergraduate biology students in single classes at one university, they demonstrated that both of these analytical tools are capable of generating assessment scores equal in precision to those by trained, expert raters (biologists with PhDs). Overall, Nehm et al. (in press) suggested that when clear scoring rubrics have been developed, and student ideas on a particular topic are well established, SIDE is much more powerful and cost effective than SPSSTA (Haudek et al., 2011; Nehm et al., in press). Since both of these factors apply to our present study, we chose SIDE as our CAS tool. For biologists who have not developed robust grading rubrics, or who have not comprehensively investigated student thinking about a topic, SPSSTA will be a more appropriate starting point (Haudek et al., 2011).

Prior studies of SIDE did not investigate several questions that arise when biologists apply scoring models beyond a single instructor, course, or college. First, are scoring models generalizable across colleges and courses (e.g., major vs. nonmajor)? That is, will successful scoring models built at one university work at another? Second, how much human scoring is needed to build a robust scoring model, and can human scoring of additional student responses compensate for scoring-model limitations across courses and colleges? Finally, what factors might account for scoring models that function effectively in a class at one university but fail at a similar class in another? Can these failures be fixed?

It is important to emphasize that CAS tools—including machine learning—are not capable of comprehending the meanings of students’ lexical responses. Programs such as SIDE simply note the presence or absence of particular words (or word pairs) in response corpora, build large matrices of word combinations, and apply sophisticated machine-learning algorithms to predict human-scoring patterns (Mayfield and Rosé, 2010a,b). Consequently, machine-learning tools are very sensitive to language, but not its meaning(s). Expert human raters, in contrast, can effortlessly comprehend diverse linguistic expressions and understand their equivalence (e.g., “some live and some die” is equivalent to “differential survival”); in contrast, computers view different text as indicative of different information. For this reason, mundane text differences, such as spelling (color vs. colour; fecundity vs. fecundity [sic]) impact scoring-model success.

Depending upon the scientific key concept for which a scoring model is built (e.g., variation, differential survival, etc.), different lexical expressions are used in different frequencies. Indeed, different populations of students—such as biology majors and nonmajors—may use characteristically different linguistic expressions to represent biological concepts. Some word combinations in some samples will be more diagnostic and predictive of key concepts than in others. Because of these concept-specific and sample-specific issues, we discuss our specific results relating to sample source (university; majors vs. nonmajors) and sample size (500 responses vs. 1056 responses) separately for each concept for which a scoring model was developed: variation, heredity, competition, and differential survival.

For key concept 1, variation, we found that SIDE scoring-model success was not sensitive to sample source (Figure 1). That is, regardless of which response corpus was used to train SIDE (i.e., OSU vs. MSU students; majors vs. nonmajors), the scoring models generated excellent agreement with trained expert raters and near-perfect kappa values (> 0.80). However, we did find that scoring models for variation were...
somewhat sensitive to sample size (that is, whether 500 or 1056 responses were used to build the scoring models). In comparison with the key concept of heredity, for example, in which a doubling of the training-sample size had almost no impact upon kappa values (adding 0.04 to 0.05), a doubling of the sample size for variation produced meaningful increases in kappa values (adding 0.14 to 0.80).

The explanation for the increase in kappa values with increasing training sample size for variation (but not heredity) appears to be related to the diversity and frequency of linguistic expressions that students used to represent these biological concepts. Although the most common term used by students to represent variation was “mutation,” various other terms were also used, such as “different alleles,” “genetic change,” or “error in DNA.” If only a few students used particular written expressions when linguistically representing the concept of variation (such as “genetic makeup”), then such expressions would be unlikely to be included in the machine-learning model, and downstream disagreements between human and computer scores would result. The frequencies of particular expressions, and their associations with other terms, influence scoring-model success. Indeed, we found that doubling the training sample for variation increased the frequencies of particular terms to a threshold at which they were included in the scoring models, producing improved kappa agreement statistics. For example, the matrix included 268 words for the n = 500 sample, while the matrix included 386 words for the n = 1056 sample. Matrix size is associated with differences in scoring-model performances.

For the concept of heredity, computer-scoring success was very stable and very successful regardless of sample size or source (Figures 1 and 3). Biology majors and nonmajors from different colleges and classes appear to use a consistent and detectable array of linguistic expressions to represent heredity concepts (e.g., Table 1).

The third concept we investigated was competition. Unlike the previous results, computer-scoring success for competition was sensitive to both sample source and sample size (Figures 1 and 3). Given our findings for variation and heredity, this result is surprising; the Nehm et al. (2010a) scoring rubric indicates that a very small set of terms is typically used to detect competition (e.g., compete, competition, competes). When we examine the frequency of students who used this concept, we find that only 1–2% of students used competition in their explanations of evolutionary change. Indeed, only 10 to 20 responses (out of 1000) included linguistic expressions relating to competition. Statistically, the probability that the algorithm will include such rare occurrences is low. Two solutions may be used to tackle the problem of rare responses: first, to amass a larger corpus of responses; or second, to use a special function in SIDE that allows users to augment the model and weight particular terms (see Nehm et al. [in press] for details). It is difficult for SIDE to build scoring models for extremely rare concepts.

The next concept we studied was limited resources. We found that the scoring models for this concept were stable in relation to both sample source and sample size (Figure 3). Kappa values were near-perfect (> 0.80) for the small data sets (n = 500) across samples, although there were some minor deviations. Overall, regardless of course or college, it appears that students commonly use consistent language patterns to represent this evolutionary concept, and scoring models for this concept work very well.

The final concept that we studied was differential survival. Similar to our findings for competition, differential survival was sensitive to both sample source and sample size. The comparatively weak performance of the differential survival scoring models was not a result of low response frequencies (as we observed with competition); large percentages of students utilized this idea in their explanations of evolutionary change (e.g., 60.3%; Figure 2). Scoring problems in this case were a product of students’ highly variable language use. This is in line with the scoring rubrics of Nehm et al. (2010a), which were built using different student samples and also note the diverse expressions with which students represent this evolutionary concept (e.g., “increase their survival,” “survived better,” “the species dies while others survive”). Because we also found that SIDE-scoring models were sensitive to sample source, different linguistic expressions may have been related to instructor discourse patterns. If, for example, students are imitating instructors’ language (cf. Nehm et al. [2010b]), and different instructors use different phrases to represent biological ideas, then the sample source will impact scoring-model efficacy (as we found). Although the scoring model built using the largest sample (n = 1056) demonstrated relatively good kappa values (e.g., 0.69, 0.89; see Figure 3), the highly variable ways of communicating the concept of differential survival appears to have limited scoring-model performance.

Generalizing Our Findings to Other Samples and Populations

Very few studies in biology education have examined the similarities and differences between different student populations’ short-answer explanations of biological phenomena, including evolutionary change. In two studies of primarily underrepresented biology students (many of whom were English-language learners) from a minority-serving institution in the eastern United States, Nehm and Reilly (2007) and Nehm and Schonfeld (2008) used short-answer, constructed-response assessments similar to those in the present study to reveal students’ thinking patterns regarding evolutionary concepts. Nehm and Schonfeld (2008) reported that their findings were generally similar to those of primarily white student populations documented in the literature. Our current findings—from primarily white, midwestern undergraduates in large, public, research universities—are also very similar to those documented in these prior studies (see Table 1). This suggests that undergraduate biology students, regardless of racial and ethnic background, may utilize a large but relatively constrained set of concepts when conceptualizing evolutionary change. Nevertheless, such conjecture should be tested using diverse student samples from different geographic regions of the country. Until such work is completed, we cannot with confidence argue that machine-learning tools will be effective for assessing all introductory biology students.

Implications for Introductory Biology Faculty

Our study has produced robust, automated, and generalizable scoring models capable of detecting most (but not all) of
the core evolutionary concepts emphasized in standards documents, curricula, and textbooks (Nehm et al., 2009b). Biology educators can make use of our work by downloading the free software package SIDE (see Mayfield and Rosé, 2010a,b) and incorporating our scoring models (freely available from the senior authors) to evaluate their students’ written explanations of evolutionary change. Using a PC computer with an i7 processor, scoring 1000 written responses takes seconds to a few minutes (depending on the concept) and produces high levels of accuracy that are comparable with consensus scores generated by two trained biologists (see Figure 4).

In addition to a user’s manual (Mayfield and Rosé, 2010a), and details on the workings of SIDE (Mayfield and Rosé, 2010b), learning how to use SIDE is illustrated in a series of video tutorials (freely available at http://evolutionassessment.org). Given that this emerging form of assessment research is new, it is important to emphasize that the software is not packaged in a user-friendly format, and like other technological tools (e.g., clickers, new operating systems, new software), effort is required to learn to use it.

Our research to date has only validated a small set of biological concepts, and introductory biology instructors are likely to want to know how their students interpret a broader array of concepts in evolution (and other content areas). We are continuing to build scoring models for other concepts, such as naïve ideas or misconceptions of evolution (Ha and Nehm, unpublished results). We speculate that improved technology and advanced research on machine-learning assessment will enable more and more concepts to be detected in students’ written responses.

National partnerships among introductory biology educators could make future work on machine learning more efficient and cost effective. Indeed, all biology educators, regardless of whether they view automated scoring as beneficial or not, could help move the field forward by collecting large corpora of students’ written responses to different prompts across subject areas (genetics, matter and energy transformations, cell biology; Haudek et al., 2011); this would help those researchers interested in using and refining machine-learning methods. Additionally, faculty from minority-serving institutions, or those teaching large English language–learning populations, are needed to expand our knowledge base on how scientific language is used to communicate core concepts in biology.

Perhaps the most significant implication of our work for introductory biology educators is that evaluating students’ written work, especially in large classes, is not impossible. This is significant from an assessment standpoint, as we contend that the process of asking students to communicate their understanding of scientific phenomena is a worthwhile activity, regardless of whether automated methods will be employed to assess these responses (e.g., Chi et al. [1994]). When analyzing students’ written responses, we have been surprised by students’ limited capacity to communicate and explain core scientific concepts (such as evolution)—particularly those students who perform admirably on MC assessments (cf. Nehm and Schonfeld [2008]). Without providing students practice and feedback in communicating their scientific understanding, we cannot expect this situation to improve.

Future work is needed to expand our concept of what constitutes a sound explanation of evolutionary change. Quantifying students’ use of necessary and sufficient scientific elements (key concepts) as a benchmark for competency, as we have done, captures only one facet of short-answer scientific explanations (cf. Braaten and Windschitl [2011]). Logic, persuasion, and argumentation skills are also important dimensions of scientific explanation, but they were not investigated in our study. Expanding our assessment framework will likely stimulate discussions about what facets of scientific explanation are most important for fostering scientific literacy.

Implications for Biology Education Researchers

Research in the use of machine learning (and text analysis in general) in biology education is only beginning (Haudek et al., 2011; Nehm and Haertig, in press; Nehm et al., in press); much remains to be learned. A community of practice on text analysis in STEM education has recently been established (see Haudek et al. [2011] and http://aacr.crcstl.msu.edu), providing a forum for researchers interested in learning more about these innovative assessment methods. Our current study has uncovered several findings likely to be of interest to researchers motivated to pursue this line of work.

First, even though we collected large response corpora, some concepts were nevertheless quite rare, limiting model performance. A large sample (n = 500) does not guarantee high concept frequency. In many instances, we were surprised by which concepts were used by students (and which were not). Second, we documented several factors that caused problems for machine-learning methods (e.g., misspellings; linguistic diversity) that nevertheless can be addressed by using a spell-checker during data gathering and weighting text expressions prior to analysis. Third, the diversity of linguistic expressions associated with concepts was highly variable (and generally unpredictable a priori), impacting scoring success. Some concepts were easily detected by the software, whereas others were not. Overall, the process of building automated scoring models is effortful and requires clear scoring rubrics and thousands of carefully evaluated responses.

ACKNOWLEDGMENTS

We thank Kristen Smock and Judith Ridgway, OSU, and Donna Koslosky and Tammy Long, MSU, for assistance with data gathering; Luanna Prevoist for comments on the manuscript; and the Automated Analysis of Constructed Response (AACR) group for discussions. R.H.N. thanks Elijah Mayfield and Caroline Rose and the faculty of the Carnegie Mellon Pittsburgh Science of Learning Center summer school for help with machine-learning methods. We also thank two reviewers for helpful suggestions for improving the manuscript. We thank the National Science Foundation (NSF; REESE grant 090999 to principal investigator [PI] R.H.N.) for funding M.H. and collaborative NSF CCLI 1022653 to PI’s Jenny Knight, R.H.N., and M.U.-L for support of the AACR group. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the NSF. The research was conducted under OSU IRB Protocol # 2008B0080 (R.H.N., PI).
REFERENCES


Evans EM, Lane JD (2011). Contradictory or complementary? Creationist and evolutionist explanations of the origin(s) of species. Hum Dev 54, 144–159.


M. Ha et al.
presented at the National Association of Research in Science Teaching, Orlando, FL, April 2011.


Using Comparative Genomics for Inquiry-Based Learning to Dissect Virulence of \textit{Escherichia coli} O157:H7 and \textit{Yersinia pestis}

David J. Baumler,*‡ Lois M. Banta,‡ Kai F. Hung,§ Jodi A. Schwarz,|| Eric L. Cabot,† Jeremy D. Glasner,† and Nicole T. Perna†¶

*BACTER Institute and †Genome Center of Wisconsin, University of Wisconsin, Madison, WI 53706; ‡Department of Biology, Williams College, Williamstown, MA 01267; §Department of Biology, Eastern Illinois University, Charleston, IL 61920; ||Department of Biology, Vassar College, Poughkeepsie, NY 12604; ¶Department of Genetics, University of Wisconsin, Madison, WI 53706

Submitted April 12, 2010; Revised October 20, 2011; Accepted October 26, 2011

Monitoring Editor: Erin L. Dolan

Genomics and bioinformatics are topics of increasing interest in undergraduate biological science curricula. Many existing exercises focus on gene annotation and analysis of a single genome. In this paper, we present two educational modules designed to enable students to learn and apply fundamental concepts in comparative genomics using examples related to bacterial pathogenesis. Students first examine alignments of genomes of \textit{Escherichia coli} O157:H7 strains isolated from three food-poisoning outbreaks using the multiple-genome alignment tool Mauve. Students investigate conservation of virulence factors using the Mauve viewer and by browsing annotations available at the A Systematic Annotation Package for Community Analysis of Genomes database. In the second module, students use an alignment of five \textit{Yersinia pestis} genomes to analyze single-nucleotide polymorphisms of three genes to classify strains into biovar groups. Students are then given sequences of bacterial DNA amplified from the teeth of corpses from the first and second pandemics of the bubonic plague and asked to classify these new samples. Learning-assessment results reveal student improvement in self-efficacy and content knowledge, as well as students’ ability to use BLAST to identify genomic islands and conduct analyses of virulence factors from \textit{E. coli} O157:H7 or \textit{Y. pestis}. Each of these educational modules offers educators new ready-to-implement resources for integrating comparative genomic topics into their curricula.

LEARNING OBJECTIVES

At the completion of these activities, students will:

1. have improved their ability to use BLAST;
2. be able to identify genomic islands from whole-genome alignments;
3. know one way to explore existing annotation for genes in a genomic island and determine whether any are involved in virulence;
4. be able to conduct analyses addressing conservation of virulence factors in \textit{Escherichia coli} O157:H7 or \textit{Yersinia pestis} strains; and
5. be able to apply these newly acquired skills to design a bioinformatic investigation of an \textit{E. coli} outbreak.

INTRODUCTION

The advent of genome sequencing and increased use of computational biology for analysis have dramatically changed the landscape for undergraduate student learning in the areas of genetics, molecular biology, and even ecology. A variety of curricular approaches attempt to help
students understand the impact that genomics will have in their lives and careers (Campbell, 2002; Dexter Dyer and LeBlanc, 2002; Honts, 2003; Petril and Justice, 2007). A number of efforts are currently underway nationwide to develop new teaching approaches and resources for undergraduate curricula in bioinformatics and genomics (Forst and Goodner, 2006; Goodner and Wheeler, 2006; Baumler et al., 2008; Lopatto et al., 2008; Lopatto and Elgin, 2010; Shaffer et al., 2010; the Joint Genome Institute’s “adopt a microbial genome” [www.jgi.doe.gov/education/genomeannotation.html]; the HHMI–Scientific Education Alliance’s phage annotation initiative [www.hhmi.org/grants/sea]; and the Teaching Big Science at Small Colleges: A Genomics Collaboration initiative [http://serc.carleton.edu/genomics/index.html]). These pioneering educational initiatives reaffirm and build on the premise that learning objectives are met and exceeded when students find the topics engaging, exciting, and worthwhile (National Research Council, 2005). Most of these teaching resources address student learning about genomics and bioinformatics through active or inquiry-based learning (Gallagher et al., 1995; Checkley, 1997). Active learning of these topics requires the use of computers to develop and reinforce skills in bioinformatic gene analysis. Microbial genomes are especially suitable for teaching these subject areas, since they are relatively small, have modest computational requirements, and present a way to initiate student learning by investigation and problem solving of real-world, ill-defined problems (Wiggins and McTighe, 2005; Nehm, 2010). There are a variety of existing educational resources that focus on a single genome (e.g., Lopatto et al., 2008), and others designed to teach about genomic technologies (e.g., Shuster, 2011). However, despite the recent advances in DNA-sequencing technology that have resulted in an abundance of available genome sequences, relatively few curricular modules address genome-level questions. As of 2011, more than 1000 microbial genomes have been sequenced, representing a vast untapped resource for educators interested in using inquiry-based exercises that can compare multiple genomes from different strains of related microbes.

Comparative genomics is the study of the relationships of genome content, structure, and function across multiple organisms. This field is growing rapidly alongside advances in sequencing, particularly for studies of very closely related species and strains of bacteria. Comparative genomic techniques are used to discover genetic variation at scales ranging from large-scale chromosomal rearrangements to single-nucleotide polymorphisms or SNPs. The presence or absence of protein-coding genes (open reading frames or ORFs) often leads to new hypotheses regarding the distinguishing traits of each microorganism. While a number of comparative genomic tools exist, many are limited to the analysis of two genomes. One tool that permits the comparison of more than two genomes is the multiple-genome alignment tool Mauve (Darling et al., 2004). Mauve is particularly appealing as an educational tool, because it includes a user-friendly visualization of the results, and runs on widely available, modestly priced computer hardware. Mauve has been used extensively for comparative genomic analysis of numerous types of pathogenic microorganisms, such as Pseudomonas spp. (Glasner et al., 2008), Salmonella spp. (Vernikos et al., 2007), E. coli (Mau et al., 2006), and Y. pestis (Darling et al., 2008). These studies illustrate the applicability of comparative genomics for deciphering and identifying unique genomic regions that may play a role in the strain-to-strain variation of pathogenic microorganisms.

Key signatures of pathogenesis-promoting regions are genes encoding virulence factors or proteins that play a role in the ability of the pathogen to cause disease. Some types of microbial virulence factors are: adhesins (which help the microbe attach to host tissue), toxins (secreted proteins toxic to host cells or organs), secretion systems (which inject microbial proteins into the host cell or environment), and defenses against host barriers (which protect the microbe from the host’s immune system or, e.g., allow passage through the acidic fluid in human stomachs to reach the intestine). For pathogenic members of the bacterial family Enterobacteriaceae, which includes the human pathogens E. coli O157:H7, Salmonella, and Y. pestis, a concerted effort to identify virulence factors based on experimental evidence and/or bioinformatic analysis has been conducted, and the resulting annotations that support the designation of genes as “putative or known virulence factors” are available in the A Systematic Annotation Package for Community Analysis of Genomes (ASAP) database (Glasner et al., 2006; http://asap.aghs.wisc.edu/asap/home.php).

We have developed two curricular modules to support student learning of bioinformatic skills via investigation of engaging questions in comparative microbial genomics. The extensive existing and recently updated ASAP annotation allows us to focus students’ attention on downstream inferences, while the real-world clinical relevance appeals to the many premedical students in our courses. In the first module, students focus on three genomes of the human pathogen E. coli O157:H7 from separate foodborne outbreaks; a second module centers on student analysis of five genomes of the human pathogen Y. pestis. The skills addressed in both modules are the use of the basic local assignment search tool (BLAST; Altschul et al., 1990) and identification and analysis of genomic islands. Through these exercises, students are exposed to core concepts in microbial genomics, including functions and conservation of virulence factors; horizontal gene transfer; the evolution/structure/function of genomic islands; and SNPs (in genes unrelated to pathogenesis) as possible determinants of metabolic capabilities. In crafting the assignments for these modules, we focused on questions that allowed students to apply both their newly developed skills and these core concepts. For example, students use the literature resources linked to the ASAP database to investigate the function and potential relevance to pathogenicity of a gene on a genomic island; this task requires understanding of the concept of virulence factors and some knowledge of possible virulence factor functions, as well as the ability to use the Mauve alignment tools to identify a gene of interest. We also sought to foster higher-level thinking skills, by posing questions that require problem solving, analysis, and synthesis (Allen and Tanner, 2002).

The three E. coli O157:H7 strains in the first module were isolated from sickened individuals or contaminated food during outbreaks associated with ground beef in Michigan in 1982 (Perna et al., 2001); radish sprouts in Sakai City, Japan, in 1996 (Hayashi et al., 2001); and fresh bagged spinach in 17 states in the United States in 2006 (Manning et al., 2008). Strains of E. coli O157:H7 differ in the severity of disease
they cause in humans and can lead to bloody diarrhea, renal failure, hemolytic uremic syndrome (HUS), or death. The three strains of \textit{E. coli} O157:H7 genomes used for this exercise, strain EDL933 (ground beef), Sakai (radish sprouts), or EC4042 (bagged spinach), each have different epidemiological statistics for those sickened from the outbreaks and differed in their hospitalization-to-death ratios of 23:0 (EDL933), 8938:3 (Sakai), and 205:3 (EC4042) (Hayashi et al., 2001; Rangel et al., 2005; Manning et al., 2008). On the basis of the epidemiological statistics, students learn that each of these strains differs in the severity of the disease that they cause in humans (Manning et al., 2008). In this module, students use Mauve to perform a comparative genomic analysis, looking for the presence or absence of virulence factor genes and eventually generating bioinformatics-based hypotheses to address strain-to-strain variation in pathogenicity and epidemiological outcomes.

The \textit{Y. pestis} module reinforces the bioinformatic skills and core concepts learned in the first module, and provides an opportunity for students to delve further into analysis of genetic events, such as insertions, deletions, and SNPs responsible for interstrain variation. \textit{Y. pestis} is the causative agent responsible for three historical global pandemics of the bubonic plague or Black Death (~550 A.D., 1347 A.D., 1850 A.D.; Drancourt et al., 2004; Stenseth et al., 2008). This human pathogen is notorious for its ability to cause widespread death, as in the second pandemic, in which ~30–60% of the entire European population succumbed. Through use of the instructional materials provided, students learn about two routes of transmission for \textit{Y. pestis} infection of humans (Chamberlain, 2004). In the first, fleas transfer the bacterium by first biting infected small rodents, such as rats, and subsequently biting humans, resulting in bacterial infection in the human bloodstream (bacteremia). The second route is through human respiratory infection, in which infected individuals spread the bacteria to others via droplet infection. Students also learn that strains of \textit{Y. pestis} are typically categorized into one of three biovars (Antiqua [east African origin], Mediaevalis [central Asian origin], or Orientalis [central Asian origin]) based on their experimentally determined ability to use the carbon sources arabinose or glycerol and the nitrogen source nitrate (Devignat, 1951), and discover that the inability of \textit{Y. pestis} strains to utilize one of these two carbon sources or the nitrogen source can be traced to mutations in one of three genes, \textit{araC}, \textit{glpD}, or \textit{napA}. This gene analysis provides an opportunity for instructors to introduce or reinforce students’ content knowledge, including the connection between genetic traits and biochemical pathways; pseudogenes, missense, and nonsense codons; and the potential effects of single amino acid changes on metabolic enzyme activity. The five \textit{Y. pestis} genomes in this module (CO92 [Parkhill et al., 2001], KIM [Deng et al., 2002], 91001 [Song et al., 2004], Antiqua and Nepal [Chain et al., 2006]), include at least one representative of each biovar. Using BLAST, students classify each strain into biovars based on analysis of the three genes. BLAST comparisons with the genome of a sixth strain isolated from North America (YPE), allow students to deduce whether Pacific or Atlantic Ocean trade routes were most likely responsible for the arrival of \textit{Y. pestis} in North America. Additionally, students use BLAST comparisons of modern \textit{Y. pestis} DNA with sequences generated from the dental pulp of corpses of humans thought to have died during the first and second pandemics of the Black Death (Drancourt et al., 1998, 2004; Tran-Hung et al., 2007) to determine which biovar(s) are most similar to the strains that caused the two previous pandemics. In a final component of this module, students apply the comparative genomic techniques learned from the \textit{E. coli} module to address variation in virulence factors among \textit{Y. pestis} strains. One strain included in the genome alignment (91001) has lost the ability to cause disease in humans (Song et al., 2004). Students identify genomic islands conserved in four genomes (Antiqua, Nepal, KIM, CO92), but absent in the genome of strain 91001. Analysis of the genes contained within permits students to formulate new hypotheses about why the genes in these islands may be important for causing human disease.

To assess the effectiveness of these modules, we collected and analyzed data on students’ self-efficacy, abilities, and content knowledge. Our pre- and posttest assessment data indicate that students reported gains in their ability to use BLAST and to analyze gene content and conservation in genomic islands. Analytical skills and content knowledge–based assessments support the claim that students achieved the learning objectives and were able to apply their newly acquired abilities to address open-ended questions requiring experimental design, deductive reasoning, and literature-based analyses of experimental evidence. Throughout this work, we refer to use of the BLAST and Mauve tools as “abilities” and the application of these abilities to formulating hypotheses, problem solving, analysis, and synthesis as “skills.” These educational modules are the first to integrate epidemiological, phenotypic, paleomicrobiological, and bioinformatic analyses to help students learn about the consequences of variation in genome content among human pathogens. Moreover, they offer educators new, engaging resources to immediately integrate comparative genomic topics into their undergraduate curricula.

**MATERIALS AND METHODS**

**General Study Design**

These modules were used and learning-assessment data were collected from nine educational settings at three large public universities and two private, liberal arts colleges (Table 1). Consent was sought and granted by students in cohorts 1 and 6–9 in accordance with Institutional Review Board (IRB) guidelines for all student data presented from the use of these modules. Participation in the assessments by students in

<table>
<thead>
<tr>
<th>Sample number</th>
<th>Course</th>
<th>Year</th>
<th>Number of students</th>
<th>\textit{E. coli} module</th>
<th>\textit{Y. pestis} module</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bioinformatics</td>
<td>2009</td>
<td>13</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>2</td>
<td>Microbiology</td>
<td>2008</td>
<td>15</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>3</td>
<td>Microbiology</td>
<td>2008</td>
<td>12</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>4</td>
<td>Microbiology</td>
<td>2009</td>
<td>20</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>5</td>
<td>Microbiology</td>
<td>2010</td>
<td>23</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>6</td>
<td>Genomics</td>
<td>2009</td>
<td>10</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>7</td>
<td>Genomics</td>
<td>2010</td>
<td>17</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>8</td>
<td>Microbiology</td>
<td>2011</td>
<td>14</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>9</td>
<td>Microbiology</td>
<td>2008</td>
<td>14</td>
<td>Y</td>
<td>N</td>
</tr>
</tbody>
</table>
cohort 2–5 was voluntary, in that students could choose not to complete the assessments. The IRB boards at all institutions either approved these studies or ruled these studies exempt. Analysis of statistical significance was performed using the Student’s paired t test with a one-tailed distribution and two-sample equal variance.

**Participant Population**

Student demographic information was collected for all samples. Of the 138 students who participated in these studies, 52 were male and 86 were female. Twenty-six were sophomores, 53 were juniors, 41 were seniors, and 18 were graduate students. The racial/ethnic composition of the participant pool was: 94 Caucasian, 21 Asian, 14 Latino/Hispanic, four African American, two African, one student from the Indian subcontinent, one Native American, and one Latino/Pacific Islander. The degree majors represented among the students included biology (81%), biochemistry (4%), chemistry (4%), economics (2%), and microbiology (9%).

**Assessing Changes in Self-Efficacy**

Student self-efficacy assessment was performed as previously described (Likert, 1932) using pre- and posttest self-report questions administered to students (Table 2). Pretest data represent a compilation of the data from samples 1–8 (n = 124). Posttest data were separated for classes that used only the E. coli module (samples 2, 4, 6, 7, and 8 [n = 76]) and for those that used both modules (samples 1, 3, and 5 [n = 48]). Self-efficacy assessment questions 1, 2, and 3 were used in all but one course (samples 1 through 8), while assessment question 4 was used in only five courses (samples 1, 5, 6, 7, and 8).

**Assessing Changes in Abilities and Higher-Level Thinking Skills**

To determine whether student learning gains occurred through use of these modules, formative assessments (Hutchings, 2000; Mettetal, 2001) of students’ abilities to use their newly acquired BLAST and Mauve abilities were conducted using standardized rubrics to evaluate written responses to the course assignments (see course assignments 1 for student samples 1, 4, 8, and 9 and 2 [samples 1 and 3] in the Supplemental Material) turned in by students or pairs of students (Tables 3 and 4). Written student products also provided the data to assess whether these resources can be used to promote higher-level analysis (Table 5) and synthesis (Supplemental Figure S7) skills through guided student inquiry (samples 2 and 3).

**Assessing Changes in Content Knowledge**

To further determine whether student learning gains occurred through use of these modules, assessments of students’ content knowledge were conducted by evaluating written student responses to the course assignments (see course assignments 1 and 2 in the Supplemental Material), as described in the preceding section. Acquisition and retention of content knowledge were further assessed using qualitative evaluation of voluntary student written responses to pre- and posttest questions (sample 5; Table 6). Students’ written responses to a take-home exam question about a hypothetical

---

**Table 2. Pre- and posttest self-efficacy assessment (n = 124)**

<table>
<thead>
<tr>
<th>Student response option</th>
<th>Pretest (% ± SE)</th>
<th>Posttest E. coli only (% ± SE)</th>
<th>Posttest E. coli and Y. pestis (% ± SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question 1: I use BLAST frequently and am confident in my ability with it.</td>
<td>6.5 ± 2.7</td>
<td>22.4 ± 4.3</td>
<td>41.0 ± 12.4</td>
</tr>
<tr>
<td>Strongly disagree</td>
<td>67.6 ± 10.0</td>
<td>0 ± 0</td>
<td>0 ± 0</td>
</tr>
<tr>
<td>Disagree</td>
<td>30.4 ± 11.5</td>
<td>8.0 ± 5.2</td>
<td>16.0 ± 12.5</td>
</tr>
<tr>
<td>Agree</td>
<td>2.0 ± 1.3</td>
<td>62.6 ± 11.7</td>
<td>52.0 ± 8.1</td>
</tr>
<tr>
<td>Strongly agree</td>
<td>0 ± 0</td>
<td>29.4 ± 14.5</td>
<td>32.0 ± 20.7</td>
</tr>
<tr>
<td>Question 2: I am familiar with BLAST and could probably find my way around with it.</td>
<td>36.9 ± 18.6</td>
<td>77.6 ± 4.8</td>
<td>56.5 ± 11.4</td>
</tr>
<tr>
<td>Strongly disagree</td>
<td>22.8 ± 8.8</td>
<td>0 ± 0</td>
<td>0 ± 0</td>
</tr>
<tr>
<td>Disagree</td>
<td>6.5 ± 10.0</td>
<td>0 ± 0</td>
<td>0 ± 0</td>
</tr>
<tr>
<td>Agree</td>
<td>2.0 ± 1.3</td>
<td>62.6 ± 11.7</td>
<td>52.0 ± 8.1</td>
</tr>
<tr>
<td>Strongly agree</td>
<td>0 ± 0</td>
<td>29.4 ± 14.5</td>
<td>32.0 ± 20.7</td>
</tr>
<tr>
<td>Question 3: I have heard of BLAST and have a vague idea of what it is.</td>
<td>33.7 ± 12.2</td>
<td>0 ± 0</td>
<td>2.5 ± 2.5</td>
</tr>
<tr>
<td>Strongly disagree</td>
<td>22.8 ± 8.8</td>
<td>0 ± 0</td>
<td>0 ± 0</td>
</tr>
<tr>
<td>Disagree</td>
<td>6.5 ± 10.0</td>
<td>0 ± 0</td>
<td>0 ± 0</td>
</tr>
<tr>
<td>Agree</td>
<td>2.0 ± 1.3</td>
<td>62.6 ± 11.7</td>
<td>52.0 ± 8.1</td>
</tr>
<tr>
<td>Strongly agree</td>
<td>0 ± 0</td>
<td>29.4 ± 14.5</td>
<td>32.0 ± 20.7</td>
</tr>
<tr>
<td>Question 4: Given a gene that confers a virulence trait in one pathogen, I know how to determine whether the gene is conserved in other related pathogens.</td>
<td>31.7 ± 21.8</td>
<td>0 ± 0</td>
<td>0 ± 0</td>
</tr>
<tr>
<td>Strongly disagree</td>
<td>31.7 ± 21.8</td>
<td>0 ± 0</td>
<td>0 ± 0</td>
</tr>
<tr>
<td>Disagree</td>
<td>39.0 ± 9.2</td>
<td>2.7 ± 1.8</td>
<td>5.2 ± 5.1</td>
</tr>
<tr>
<td>Agree</td>
<td>26.8 ± 14.5</td>
<td>64.8 ± 19.7</td>
<td>53.8 ± 2.4</td>
</tr>
<tr>
<td>Strongly agree</td>
<td>2.5 ± 2.3</td>
<td>32.5 ± 20.7</td>
<td>41.0 ± 5.6</td>
</tr>
</tbody>
</table>
2. Be able to identify genomic islands from whole genome alignments
   Did the student explore the annotations for genes in a genomic island and determine whether any are involved in virulence
3. Know one way to explore existing annotation for genes in a genomic island and determine whether any are involved in virulence
   Did the student identify gene products located in a genomic island?
4. Be able to conduct analyses addressing conservation of genes in E. coli O157:H7 or Y. pestis strains
   Did the student determine using BLAST and Mauve whether the assigned virulence gene was present in other E. coli genomes?

a Items were scored as Yes/No unless otherwise noted.
b Students correctly identified a wide variety of phage-related genes (e.g., those encoding replication proteins, portal proteins, hydrolases, capsid components, and tape measure proteins), as well as transposase and integrase genes, as being unrelated to virulence. Genes identified by students as being highly relevant to differences in pathogenicity included shiga-like toxins and the host-cell adhesion gene iha (one student cited literature showing that iha allowed the bacterium to adhere to kidney cells). Other students pointed to ureases and a short-chain dehydrogenase/reductase as potentially conferring the ability to survive in new environments and/or exploit unique nutritional resources. Finally, students hypothesized that complement resistance proteins and proteins involved in resistance to oxidative stress and phagocytic activity could contribute to a strain’s virulence by enabling it to withstand host defenses.

Table 3. Scoring rubric for formative assessment of student learning based on written responses to assignment (course assignment 1 in the Supplemental Material) on E. coli module

<table>
<thead>
<tr>
<th>Learning objective</th>
<th>Did the student conduct a BLAST search of his/her assigned virulence gene?</th>
<th>Did the student determine whether the assigned gene was present in microorganisms other than E. coli?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Improve student’s ability to use BLAST</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Be able to identify genomic islands from whole genome alignments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Know one way to explore existing annotation for genes in a genomic island</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Be able to conduct analyses addressing conservation of genes in E. coli O157:H7 or Y. pestis strains</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Scoring rubric for formative and skills assessment of student learning based on written responses to assignment (course assignment 2 in the Supplemental Material) on Y. pestis module

<table>
<thead>
<tr>
<th>Learning objective</th>
<th>Did the student conduct a BLAST search of the glpD, napA, and araC genes?</th>
<th>Did the student correctly interpret the SNP and BLAST data and successfully assign each strain and dental pulp sample to the correct biovar?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Improve student’s ability to use BLAST</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Be able to identify genomic islands from whole genome alignments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Know one way to explore existing annotation for genes in a genomic island</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Be able to conduct analyses addressing conservation of genes in Y. pestis strains</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a Tasks were scored as Yes/No unless otherwise indicated.
b Students typically pinpointed genes implicated in iron uptake and metabolism as being potentially relevant to virulence, because bacteria require iron and exhibit tight regulation of these functions. Other positive findings included fimbrial proteins, which the students suggested were involved in adhesion to host cells. Students correctly inferred that putative phage tail proteins, antirepressors, host-specificity proteins, and transposases are unlikely to be involved in virulence. One student found a putative sulfatase and sulfatase modifier and concluded that these genes were insufficient to cause virulence because they were found in the nonpathogenic E. coli strain K12 as well.
Table 5. Learning objectives and observed student outcomes upon completion of E. coli O157:H7 or Y. pestis modules

<table>
<thead>
<tr>
<th>Learning objective</th>
<th>Student outcome</th>
<th>E. coli O157:H7 module</th>
<th>Y. pestis module</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Improve student’s ability to use BLAST</td>
<td>Students were able to correctly identify the gene and protein sequence of the genes possibly involved in virulence.</td>
<td>98% (47/48)</td>
<td>90% (43/48)</td>
</tr>
<tr>
<td>2. Be able to identify genomic islands from whole-genome alignments</td>
<td>Students were able to correctly identify islands of conservation or dissimilarity from whole-genome alignments.</td>
<td>100% (48/48)</td>
<td>98% (47/48)</td>
</tr>
<tr>
<td>3. Know one way to explore existing annotation for genes in a genomic island and determine whether any are involved in virulence</td>
<td>While the students were able to identify the hypothetical functions of suspected virulence genes, their conclusions about whether the genes are actually involved in virulence were tentative, because they did not feel that they understood the virulence pathways in E. coli very well.</td>
<td>98% (47/48)</td>
<td>81% (39/48)</td>
</tr>
<tr>
<td>4. Be able to conduct analyses addressing conservation of genes in E. coli O157:H7 or Y. pestis strains</td>
<td>Students were able to identify whether the genes exist in all three genomes or just one, and whether identified islands are present in different strains.</td>
<td>96% (46/48)</td>
<td>83% (40/48)</td>
</tr>
<tr>
<td></td>
<td>Students were able to correctly deduce the biovars of the five strains and the dental pulp samples using BLAST results, as well as obtaining protein and ORF information following BLAST searches. In their answers, they were able to explain what the e values mean and draw the correct conclusions.</td>
<td>100% (17/17)</td>
<td>94% (16/17)</td>
</tr>
<tr>
<td></td>
<td>Students were able to correctly identify islands of conservation or dissimilarity from whole-genome alignments.</td>
<td>100% (17/17)</td>
<td>Not applicable</td>
</tr>
<tr>
<td></td>
<td>Students were able to identify annotated functions of predicted proteins encoded on a chosen island, as well as obtain information on related proteins using InterPro Scan data. Students were able to deduce that the presence or absence of a few virulence factors may or may not be sufficient evidence to conclude whether a strain is virulent or not.</td>
<td>100% (17/17)</td>
<td>88% (15/17)</td>
</tr>
<tr>
<td></td>
<td>Students were able to determine whether their assigned virulence gene is present in all five strains, but fewer than half analyzed whether the gene was functional. Students utilized BLAST similarity comparisons to draw conclusions on whether the strains originated from the Pacific or Atlantic region.</td>
<td>100% (11/11) for gene presence; 45% (5/11) for gene function</td>
<td>71% (12/17)</td>
</tr>
</tbody>
</table>

aData are reported as percent of students who successfully accomplished the task outlined in the scoring rubrics (Tables 3 and 4).
**Table 6.** Assessment of student acquisition and retention of content knowledge

<table>
<thead>
<tr>
<th>Question and evaluation of answer</th>
<th>Pretest</th>
<th>Posttest</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is a genomic island?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No idea/no answer</td>
<td>65%</td>
<td>7%</td>
</tr>
<tr>
<td>Wrong</td>
<td>17%</td>
<td>0%</td>
</tr>
<tr>
<td>Partial credit</td>
<td>22%</td>
<td>36%</td>
</tr>
<tr>
<td>Correct</td>
<td>0%</td>
<td>43%</td>
</tr>
<tr>
<td>How does genetic information get into an island?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No idea/no answer</td>
<td>65%</td>
<td>14%</td>
</tr>
<tr>
<td>Wrong</td>
<td>13%</td>
<td>0%</td>
</tr>
<tr>
<td>Partial credit</td>
<td>22%</td>
<td>43%</td>
</tr>
<tr>
<td>Correct</td>
<td>0%</td>
<td>43%</td>
</tr>
</tbody>
</table>

Suppose you had a group of bacterial strains from different regions of the world that are all the same species, yet some are more virulent than others. If you sequenced the genomes of all these strains, what feature(s) would you look for in those genome sequences that might confer strain-to-strain variability in virulence among bacterial strains of the same species?

- Full credit was given for responses that included the concept of sequences or sets of genes unique to one strain within a bacterial species that may confer virulence.
- Partial credit responses lacked the possible link to virulence.
- Full credit responses included both some mention of modes of horizontal gene transfer (e.g., phage transduction) and evidence, such as phage gene remnants and/or transposition-related sequences (transposes, insertion sequences). Partial credit responses typically failed to answer the second question.
- Acceptable responses described possible virulence factor functions (e.g., toxins, iron uptake, adhesins, etc.). Acceptable answers invoked differences in gene content without mentioning specific potential functions. Partial credit was given for “genomic islands.”

---

**E. coli** outbreak served as an additional mode of assessing student content knowledge (sample sets 7 and 8).

**Instructional Resources**

Information about the computational requirements for using Mauve and materials, methods, and figures pertaining to the **E. coli** and **Y. pestis** genome alignments are provided for instructors (see the Supplemental Material). Introductory materials for instructors to use with their courses are available as PowerPoint slides for **E. coli** O157:H7 (Supplemental Slide S1) and **Y. pestis** (Supplemental Slide S2). These slides introduce the tools used in these modules and provide background genomic, epidemiological, and historical information. For instructors and/or students, information about the use and interpretation of BLAST results can be found in the BLAST information guide available at National Center for Biotechnology Information (www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html) and additional simple explanations about using BLAST can be found in chapter 7 of the book *Bioinformatics for Dummies* (Claverie and Notredame, 2003).

**Student Instructional Resources**

Instructions for students are provided to assist with using the Mauve alignments and the ASAP database (see instructional material 1 and 2 in the Supplemental Material). Additional independent exercises consisting of multiple inquiry-based questions that can be used with these modules are provided (see course assignments 1–3 in the Supplemental Material).

**RESULTS**

To determine the extent to which our student learning objectives were achieved, we used a variety of assessment instruments, including pre- and postmodule knowledge-based questionnaires, scoring of student responses to guided inquiry-based individual assignments, and open-ended investigations/exam questions that required students to apply their newly acquired abilities and their understanding of core concepts, such as the potential functions of virulence factors or the effects of SNPs on metabolic capabilities. In addition to these tools, which were used to assess acquisition of abilities and content knowledge, self-efficacy assessment rating scales were used to investigate whether students, through use of these problem-based learning modules, gained confidence in their abilities to use BLAST, Mauve, and the ASAP database to identify and analyze the contents of genomic islands in two microbial pathogens, and to answer interesting questions about conservation of virulence factors. In our analyses of these self-efficacy data, student responses for the posttest assessment were separated into two groups; the first student group was composed of those who solely used the **E. coli** O157:H7 module, and the second set represented those who completed both the **E. coli** O157:H7 and the **Y. pestis** modules.

**Learning Objective 1: Did Students Report Gains in Their Abilities to Use BLAST through Use of One or Both Modules?**

Self-efficacy assessment revealed that the majority of the students surveyed (56.5%) had little familiarity or confidence using BLAST prior to this exercise, and use of these modules resulted in significant (*P* < 0.05) increases in student responses that indicated familiarity (76.4% ± 4.8%) or even confidence (22.4% ± 4.4%) using BLAST upon completion of the **E. coli** module alone (Table 2). Furthermore, students who used both modules felt more confident using BLAST (41.0% ± 12.3%), in comparison with those who completed only the **E. coli** module (22.4% ± 4.4%) (Table 2). Formative assessment of the students’ abilities to conduct a BLAST search was determined using the rubrics shown in Tables 3 and 4 to score written responses handed in by students upon completion of a module. Essentially all students were able to use the Mauve interface successfully to conduct a BLAST search in **E. coli** for an assigned virulence factor (Table 5). Ninety percent of students properly conducted BLAST searches against other microbial genomes (Table 5). Finally, the **Y. pestis** module asks students to apply their skills with BLAST to determine the biovar most likely to have caused the first and second pandemics, and almost all students accomplished this task correctly. In their written responses to the assignments (provided in the Supplemental Material), students typically cited a correct interpretation of e values to support their...
conclusions. Overall, these results indicate that the module instructions are clear and provide evidence for success in helping students improve their self-efficacy and ability in using BLAST for inquiry-based research.

**Learning Objective 2: Were Students Able to Identify Genomic Islands Using Mauve?**

While multiple comparative genomic software tools exist, Mauve is unique, in that it allows more than two genomes to be aligned, has a student-friendly visual interface to interpret the genome alignment, and when using genomes obtained from the ASAP database, contains direct Web links from the genes to annotation pages, thus permitting students to survey additional information regarding the genes’ roles in pathogenesis. We sought to determine whether this comparative genomic tool could be used effectively by students to identify polymorphisms that differentiate strains within a given species. Not surprisingly, our pretest self-efficacy assessment data indicated that almost none of our students felt knowledgeable about comparing multiple genomes to identify genomic islands (regions unique to a single genome; Table 2). The posttest data revealed that 92% or 84% of students surveyed gained confidence in using a Mauve multiple-genome alignment to identify a genomic island upon completion of one or both exercises, respectively ($P < 0.0001$; Table 2). Analysis of written student responses to the assignments corroborated these data, indicating that the vast majority of students successfully identified a genomic island unique to a single genome or to a subset of the genomes (Table 5). On the basis of a pretest knowledge assessment, 96% of students in one class (sample 5) had no prior knowledge of what a genomic island is; by the end of the course, 5 wk after using the second module, 86% of the students had acquired and retained at least partial (50%) or accurate (36%) understanding of this core concept (Table 6). Thus, these educational modules based on Mauve appear to be an effective tool to enable students to learn important comparative genomic approaches and to acquire content knowledge.

**Learning Objective 3: Did Students Learn How to Determine Whether Any Genes in a Genomic Island May Be Involved in Virulence?**

Genomic comparison of *E. coli* K-12 (strain MG1655) and O157:H7 (strain EDL933) revealed that many of the differences that distinguish these *E. coli* strains may be attributed to the contents of genomic islands unique to the pathogenic strain (Perna et al., 2001). Genomic island analysis can also be applied to the genomes of multiple pathogenic strains to identify variations and develop new hypotheses that address epidemiological data associated with historical outbreaks. This learning objective addressed the insight that variations in the gene content of genomic islands may correlate with variations in pathogenesis from one strain to another. Therefore, we sought to determine whether changes occurred in students’ self-efficacy and in their ability to analyze the genes contained in genomic islands and to develop hypotheses to address strain-to-strain variability among microbial pathogens. Based on student self-reports (Table 2), it is clear that most students (80%) did not possess the knowledge to approach such a scientific problem prior to use of these modules. Posttest data demonstrated a significant ($P < 0.0005$) increase in self-reported ability to discern whether any of the genes/ORFs contained within a genomic island may be involved in virulence, upon completion of one (98%) or both (88%) modules (Table 2). It is interesting to note that 14% of the students who completed both modules still did not feel that they had learned enough to address this type of problem.

We used student responses to the course assignments (see the Supplemental Material) to independently assess students’ abilities to identify the genes contained within genomic islands and to relate their predicted functions to potential roles in pathogenicity. These data demonstrate that essentially all of the students successfully analyzed the gene contents of a genomic island by using the annotation links to ASAP, and that 80–90% were able to formulate a hypothesis as to why the presence or absence of the genomic island contents may contribute to the microorganism’s virulence (Table 5). Not unexpectedly, students who completed the *Y. pestis* module were more likely to have successfully addressed possible roles for the island-borne genes than students exposed to the roles of virulence factors for the first time during the *E. coli* module (Table 5). Pre- and posttest assessment revealed substantial gains in students’ content knowledge about the types of functions (e.g., adhesins, iron-carrying proteins, toxins, and secretion systems) characteristic of virulence factors (Table 6), consistent with the notion that exploration of possible virulence factors encoded within islands may have helped the students recall these functions.

Many of the genetic islands present in the *E. coli* and *Y. pestis* strains contain a large number of phage-related and insertion element–related genes, suggesting they originated through a phage-transduction event. Depending on how familiar they are with phage biology, students may or may not recognize these as phage genes or remnants from the annotations provided in ASAP. As students explore the genome island content, instructors can use the presence of these genes as a starting point for a discussion of horizontal gene transfer. In one class, written lab responses to the open-ended question “From your analyses, what is one mechanism that has caused divergence of these genomes?” showed that 10 out of 14 pairs of students were able to correctly deduce that phage transduction was responsible (data not shown). Interestingly, a pretest knowledge question posed to another class of students at the same institution revealed that 22% of the students had some inkling (almost certainly from an earlier genetics prerequisite course) of phage transduction as a possible explanation for the presence of genetic information within a genomic island, even though only 4% of those students could articulate a cogent definition of a genomic island before using the module (Table 6). After experience with both modules, 86% of the students gave a response that was at least partially correct.

**Learning Objective 4: Through Use of One or Both of These Modules, Did Students Improve Their Ability to Conduct Analyses of Virulence Factors in Multiple Genomes of Microbial Pathogens?**

To determine whether students could use BLAST as a tool to address the conservation of genes thought to play a role in pathogenesis, we provided students with genes identified as virulence factors by scientific experts at the ASAP.
to notice a large deletion in the
cause they relied exclusively on the SNP analysis and failed
of students in one sample drew the wrong conclusion, be-
cpecific trade routes (Tables 4 and 5). In this case, four of 11 pairs
synthesizing the data needed to determine whether the North
half of the students who used the
in other Enterobacteria?" (Tables 3 and 5). Tellingly, less than
but students were somewhat less competent at using BLAST
able moment" regarding the choice of nucleotide versus pro-
tect BLAST to address the question “Is this gene or a homolog found
in other Enterobacteria?” (Tables 3 and 5). Tellingly, less than
half of the students who used the Y. pestis module ascertained
whether their assigned virulence gene was actually predicted to
be functional in all the genomes, although all of the stu-
dents reported whether it was present, and the assignment
explicitly asked about gene product function (Table 5). Some
students using this module also exhibited difficulty in fully
synthesizing the data needed to determine whether the North
American lineage is more likely attributable to Atlantic or Pa-
cific trade routes (Tables 4 and 5). In this case, four of 11 pairs
of students in one sample drew the wrong conclusion, be-
cause they relied exclusively on the SNP analysis and failed
to notice a large deletion in the glpD BLAST analysis.

Questions of conservation lend themselves well to a “teach-
able moment” regarding the choice of nucleotide versus pro-
tein BLAST. In one group of 28 students, students were asked
to provide a written response justifying their choice of us-
ing BLASTP or BLASTN. Twelve of the 14 pairs of students
provided answers that were complete and exhibited clear
comprehension of relevant concepts, including third posi-
tion wobble. One pair gave an answer that was adequate, al-
though not thorough, while the last pair’s response invoked
introns, an informative answer, in that it revealed a miscon-
ception grounded in a basic understanding of the Central
Dogma, concerning the absence of splicing in bacteria.

Together, these data demonstrate that these modules can
be used both to teach about gene conservation and to help
students develop analytical and synthesis skills. To further
illustrate this point, one instructor used the assignment of
virulence factors as a springboard for a larger unit on type 3
secretion systems (T3SSs). The “locus of enterocyte efface-
ment” (LEE) genomic island is a virulence determinant en-
coding a number of proteins, including the components of
a T3SS, required for attachment of enteropathogenic E. coli
strains to host intestinal cells (Jerse et al., 1990). Found in
many plant and animal pathogens, T3SSs function to deliver
bacterial virulence factors that typically subvert host defenses
and/or manipulate host processes, such as cytoskeletal rear-
rangements (Galan and Collmer, 1999). The LEE T3SS translo-
cates substrates that cause the formation of a raised pedestal
or “docking platform” on the surface of host cells and the Tir
protein that serves as a receptor for the bacterial membrane
adhesion intimin (Kenny et al., 1997). Prior to any class dis-
cussion on T3SSs, students explored the conservation of the
LEE genes as part of the E. coli module. As one aspect of this
analysis, the students were asked to investigate the litera-
ture links for their assigned LEE gene in the ASAP database,
and to report in writing “What do we know?” (evidence that
this gene contributes to virulence); “What do we think we
know?” (predicted or known function of the encoded gene
product); and “What do we need to know more about?”
(open/unanswered questions about the protein, but also
background concepts or elements with which the students
were unfamiliar). Even in the absence of any prior knowl-
edge about T3SSs, the morphological changes incited by
the pathogen, or the key delivered substrate Tir, pairs of
students were able to identify each of these pieces of in-
formation for their assigned gene (Table 7). To gain prac-
tice in synthesizing disparate pieces of data, the students
were then organized into groups around related gene prod-
ucts and asked to assemble a concept map, with each stu-
dent “expert” contributing information on his/her individual
virulence factor (see course assignment 3 in the Supplemental
Material). The exercise culminated with the entire class as-
sembling one large, and highly accurate, concept map (Sup-
plemental Figure S7), demonstrating that students can use
the information in the ASAP database, together with litera-
ture searches, as a tool for independent analysis and synthesis
of information about a complex regulatory system.

Learning Objective 5: Could Students Apply These
Newly Acquired Skills to Design a Bioinformatic
Investigation of an E. coli Outbreak?
To determine whether students could apply the combination
of newly acquired skills and content knowledge, we asked
them to design a bioinformatic study to approach the chal-
lenge of determining which genes might be contributing to
the unique virulence of a new outbreak strain. Student
responses were evaluated on five criteria that encompassed bi-
ology, bioinformatics, and experimental process, as outlined
in Table 8. Sixty-six percent of students provided responses
that demonstrated understanding of virulence factors /genes,
and 61% of students could define the evolution, structure, and
function of pathogenicity islands (Table 8). Sixty-nine percent
of students recognized the need to examine genomic islands,
72% of students were able to describe a bioinformatic ap-
proach to identify genomic islands, and 50% of students were
able to describe how to determine the function of genes con-
tained in these islands (Table 8). Although students did not
conduct wet-lab experiments as part of these modules, 39% of
students understood that experimental verification is needed
to further demonstrate that one or more of the genes con-
tained within an island might function as virulence factors
(Table 8).

Overall, these data indicate that the student gains in abili-
ties and content knowledge aligned with student self-reports
of increased self-efficacy relevant to the five learning objec-
tives for these comparative genomic educational modules.
Both learning assessments and observed student outcomes
Table 7. Summary of student investigations of genes contained within the LEE island

<table>
<thead>
<tr>
<th>Assigned virulence gene on LEE island</th>
<th>Evidence for role in virulence</th>
<th>Known or predicted function</th>
<th>Open/unanswered questions&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Unfamiliar concepts/elements&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>escN</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>escJ</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>rOrf1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>espZ</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>cesT</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>cesD</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>espG</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>espH&lt;sup&gt;d&lt;/sup&gt;</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>espH&lt;sup&gt;d&lt;/sup&gt;</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>espF</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>sepL</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>grlA</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>grrR</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>An X indicates students successfully identified the information indicated for their assigned virulence gene.

<sup>b</sup>Answers included the need for additional experiments to confirm reported protein–protein interactions and how those interactions affect virulence, functions of gene products reported to be regulated by the assigned LEE-encoded protein, mechanisms of action of the LEE-encoded protein, and how the protein in question can act as a chaperone if it does not directly interact with its target.

<sup>c</sup>Answers included definitions of terms, including brush border remodeling, membrane ruffling, T3SS, and attachment/effacing lesions.

<sup>d</sup>Two pairs of students investigated this gene.

---

Table 8. Student responses to the following question: “Imagine you are a genomicist working for CDC and there has been an outbreak of enterohemorrhagic disease that has resulted in illness of 100,000 people and deaths of 1000 of those patients, making this the most deadly outbreak of *E. coli* that has ever been reported. Bacteria cultured from the fecal material of some of the patients all revealed the same strain of *E. coli* that had never before been described. You have funds to sequence the genome of this strain. Design a bioinformatic study in which you approach the challenge of determining which genes might be contributing to the extreme virulence of this strain.”

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Student response</th>
<th>% (n = 36)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>Student can define and describe the concept of a virulence factor, and briefly describe at least two classes of genes that function as virulence factors.</td>
<td>Provided a complete, well-reasoned answer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Provided an incomplete/superficial answer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Did not address this topic</td>
</tr>
<tr>
<td>2.</td>
<td>Student can define the evolution, structure, and function of pathogenicity islands.</td>
<td>Provided an incomplete/superficial answer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Did not address this topic</td>
</tr>
<tr>
<td>Bioinformatics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>Student can explain that a comparison of genome sequences between closely related strains of <em>E. coli</em> will reveal genomic islands unique to one or more strains, and these may represent pathogenicity islands.</td>
<td>Provided a complete, well-reasoned answer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Provided an incomplete/superficial answer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Did not address this topic</td>
</tr>
<tr>
<td>4.</td>
<td>Student can describe a bioinformatics approach to identify genomic islands.</td>
<td>Provided a complete, well-reasoned answer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Provided an incomplete/superficial answer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Did not address this topic</td>
</tr>
<tr>
<td>5.</td>
<td>Student can describe a bioinformatics approach to learn about the potential functions of genes located within the genomic island.</td>
<td>Provided a complete, well-reasoned answer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Provided an incomplete/superficial answer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Did not address this topic</td>
</tr>
<tr>
<td>Experimental process</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>Student understands that experimental verification is needed to demonstrate that one or more of the genes contained within an island might function as virulence factors.</td>
<td>Provided a complete, well-reasoned answer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Provided an incomplete/superficial answer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Did not address this topic</td>
</tr>
</tbody>
</table>
illustrate that topics of bioinformatics, comparative genomics, and virulence factor analysis are appropriate subjects for students to study to learn about these new computational inquiry-based approaches to address scientific questions.

DISCUSSION
Comparative genomics offers a unique scientific approach that enables student exploration of the relationship between evolutionary variations in genomes and epidemiological outcomes. In an effort to help update undergraduate education in the fields of genomics and bioinformatics, we designed two modules that reinforce the use of BLAST and introduce comparative genomic alignment techniques to identify strain-to-strain differences in bacterial genomes of microorganisms that cause human disease. These tools enable students to perform these analyses in much the same way as real-world scientific researchers do. Learning assessment was conducted using pre- and posttest student questionnaires to determine whether a set of five learning objectives were achieved and to examine gains in students’ self-efficacy, skills, and content knowledge in approaching these challenging tasks. In this paper, we discuss the learning objectives, results, and some of the changes we made to improve these modules based on classroom use and student feedback.

To address the first and fourth learning objectives, students used BLAST to analyze virulence factors and their conservation among genomes. Assessment data showed that using one or both of these modules improved students’ confidence in using and competence with this type of analysis. Since the initial publication by Altschul et al. (1990) describing the BLAST algorithm, the tool has been referred to as “one of the most widely used bioinformatics programs,” and this immense use is evidenced in the sheer number (>30,000) of citations of this publication in the scientific literature. This widespread use of BLAST in the scientific community argues for the importance of students gaining familiarity and confidence in using the tool. Familiarity with BLAST is also crucial for an entry-level understanding about gene annotations and deriving information about new genes. Such information is commonly based on prior results from BLAST comparisons with known genes/proteins (Baumler et al., 2008; the Joint Genomic Institute’s “adopt a microbial genome” program; and the HHMI–Scientific Education Alliance’s phage annotation initiative). Our student learning-assessment data revealed that, upon completion of one or both of these educational modules, 100% of students are familiar with and feel they can use BLAST to work/solve future problems and have gained confidence in their abilities to apply this technique and the associated content knowledge to future scientific inquiries. While this self-reported competence is certainly an overstatement of skill level, it is a good indicator of improvement through use of the exercise. Additional assessment tools corroborated these gains by documenting student skill levels (Table 5) and increases in content knowledge (Tables 6 and 8). Unfortunately, our ability to measure increases in acquisition and retention of content knowledge (Table 6) was partially compromised by the low posttest response rate, reflecting the fact that students in this cohort were under no obligation to fill out the assessment surveys. A tendency among students to provide short and less than complete answers in the rush of the end of a semester has previously been noted (Harris et al., 2009). In our case, it is entirely possible that students who were unsure of the answers may have chosen not to take the time to respond, and/or that students who had a clearer grasp of the material rushed to finish and wrote incomplete answers. This is not the case for the data reported in Table 8, in which students responded to an exam question, rather than to a voluntary survey.

In our pilot implementations in three educational contexts, the use of both educational modules, as compared with just the E. coli module, generally resulted in increased student confidence for all of our learning objectives. While these data are encouraging, it appears that there were some students who, upon completion of the second (Y. pestis) module, became less confident in determining whether genomic island contents contribute to virulence. This may be due a larger number of genes with candidate virulence factor annotations in E. coli O157:H7 (n = 394) than Y. pestis (n = 148). Thus, students have a higher probability of locating genes thought to encode virulence factors (ORFs encoding putative or known virulence factors/total ORFs) when surveying genomes of E. coli O157:H7 (7.7%) in comparison with Y. pestis (3.6%). Additionally, the second module focuses on genomic island analysis by asking students to analyze those regions apparently lost in the genome of the 91001 strain, and to speculate on which genes may account for this strain’s inability to cause disease in humans. The genes responsible for this phenomenon are currently unknown, and the solution is not straightforward; therefore, students may have felt that they did not examine the contents of the genomic islands correctly. This point notwithstanding, when learning-assessment results were compared for students who used both modules versus the single module, it appeared that the use of the second module generally reinforced gains in self-efficacy and ability (Tables 2 and 5). This increase was likely achieved by posing a second intriguing set of scientific problems, in which the same tool can be used to further analyze similarities and differences of gene content among microbial pathogens.

Once students gained confidence in comparing single genes using BLAST, we sought to scaffold skills by providing students with hands-on experience in utilizing genome-scale BLAST queries, stored and readily available in Web-accessible databases, such as ASAP. One distinguishing feature of the ASAP database in comparison with other microbial genomics resources is that ASAP contains a copious amount of information added as “annotations.” These annotations provide standard information, such as what the gene product/protein is (product annotation) and what the protein does (function), along with many other annotation types not commonly found in other genomic resources, including mutant phenotypes, curator comments, molecular interactions, over-expression phenotypes, and virulence factor classifications. Another distinct feature of the ASAP database is a clickable link alongside each annotation that directs the user to the source of evidence from which the annotation was derived, such as a scientific publication indexed in PubMed or other bioinformatic resource(s). In the case of pathogenic microorganisms, such as E. coli O157:H7 or Y. pestis, some genes contain numerous annotations supporting the notion that the gene product may contribute to the organism’s ability to cause human disease.
This extensive collection of annotations for virulence factor genes serves as a powerful resource for students as they analyze the gene contents of genomic islands. In the E. coli O157:H7 module, students are asked to determine whether any of the genes in an island are similar to those for which evidence exists supporting their role in pathogenesis. In the Yersinia module, using a combination of BLAST, ASAP, and Mauve enables students to analyze genotypic differences across multiple microbial genomes and derive bioinformatics-driven hypotheses that may address known epidemiological, historical, experimental, and phenotypic information. The extended LEE unit we have described allowed students to discover for themselves, in a cooperative learning situation, what is known about one well-characterized virulence determinant. Among the outcomes demonstrated by the students who worked on this unit were successful evaluation of gaps in their own knowledge, ability to formulate questions (Table 7), and the ability to synthesize a complex set of information (Supplemental Figure S7); these are abilities associated with critical thinking and intellectual maturity (Allen and Tanner, 2002; Beck et al., 2010). We offer this unit as a model for instructors who wish to leverage the well-documented advantages of student-driven inquiry over conventional lectures (e.g., Knight and Wood, 2005; Armbruster et al., 2009) with the rich resource provided by the ASAP annotations.

In our experience, students tend to learn more when intrigued by problems or when questions with no “correct” answer are provided, and we used this pedagogical strategy to motivate student learning. Our approach, in which students use BLAST to address unresolved research questions and develop hypotheses about strain-to-strain differences in virulence, pathogenicity, epidemiology, and paleomicrobiology is integrative and encourages students to achieve the learning objectives of this exercise by engaging them in real-world problems. In addition to teaching about cross-genome comparisons, these modules expose students to information about a variety of virulence factors and the mechanisms by which they are thought to play a role in pathogenicity. The additional supplemental information provided as a component of these modules represents a comprehensive up-to-date survey of experimental literature related to genes in E. coli O157:H7 or Y. pestis. The different types of virulence factors are subcategorized and briefly described to provide students and instructors requisite knowledge about microbial virulence factors. This resource offers may possible avenues for expanded curricula based on these starting exercises.

Finally, one more unique and powerful feature of this set of instructional modules is that it can be used in a broad range of curricular settings. In this study, these modules were used in upper-level microbiology courses to help students learn how to study virulence, but they were also used in a genomics course to allow students to see how genomes evolve and confer new traits on organisms, and in a bioinformatics course. This set of instructional materials would likewise be entirely suitable for courses in other diverse areas of biology, including evolution or epidemiology, and could be combined with a set of wet-lab activities in which students transform E. coli, for example, to teach students about biology, not just about bioinformatics. In sum, these tools can serve as a vehicle to foster student understanding of bacterial biology, concepts of virulence and disease, proteins and genes, evolutionary mechanisms, how to work with genomic data, and how to use bioinformatics tools, potentially all in 1 wk.

CONCLUSION

Overall, these two modules represent a novel pedagogical approach to teaching molecular microbiology and comparative genomics; they focus on real-life, human disease–related questions that motivate students to learn and use bioinformatics by employing the same tools and resources used by researchers in the field. Learning-assessment results demonstrate that significant student gains in self-efficacy, ability, and content knowledge were achieved in 1) using BLAST; 2) understanding how to identify genomic regions of interest from a multiple genomic alignment; 3) analyzing the contents of genomic islands to derive bioinformatics-driven hypotheses relating to strain differences; and 4) learning about gene conservation. Our data further indicate that the depth of learning for most students is greater after using both modules. These educational modules are the first to integrate epidemiological, phenotypic, paleomicrobiological, and bioinformatics questions to help students learn about the consequences of variation in genome content among human pathogens. They offer instructors in a range of disciplines engaging new pedagogical resources for integrating comparative genomic topics into undergraduate curricula and have been implemented, tested, and refined in a variety of course contexts.

ACKNOWLEDGMENTS

This work was funded by the United States Department of Energy Genomics:GTL and SciDAC Programs (DE-FG02-04ER25627) for the BACTER Institute Post-Doctoral Fellowship (D.J.B); the National Library of Medicine, National Institutes of Health (NIH), grant number 5T15LM007359 to the Computation and Informatics in Biology and Medicine Training Program for a Post-Doctoral Traineeship (D.J.B); and in part with federal funds from the National Institute of Allergy and Infectious Diseases, NIH, Department of Health and Human Services, under contract number HHSN266200400040C. We thank Dr. Madeline Fisher for writing assistance and Drs. Bradley Anderson, Bob Mau, and Aaron Darling for technical assistance with this work.

REFERENCES


Song Y et al. (2004). Complete genome of Yersinia pestis strain 91001, an isolate avirulent to humans. DNA Res 11, 179–197.


Learning science requires higher-level (critical) thinking skills that need to be practiced in science classes. This study tested the effect of exam format on critical-thinking skills. Multiple-choice (MC) testing is common in introductory science courses, and students in these classes tend to associate memorization with MC questions and may not see the need to modify their study strategies for critical thinking, because the MC exam format has not changed. To test the effect of exam format, I used two sections of an introductory biology class. One section was assessed with exams in the traditional MC format, the other section was assessed with both MC and constructed-response (CR) questions. The mixed exam format was correlated with significantly more cognitively active study behaviors and a significantly better performance on the cumulative final exam (after accounting for grade point average and gender). There was also less gender-bias in the CR answers. This suggests that the MC-only exam format indeed hinders critical thinking in introductory science classes. Introducing CR questions encouraged students to learn more and to be better critical thinkers and reduced gender bias. However, student resistance increased as students adjusted their perceptions of their own critical-thinking abilities.

INTRODUCTION

Higher-level processing of learned information, or critical thinking, is generally viewed as an essential part of college training (American Association for the Advancement of Science [AAAS], 1990, 1993, 2010; Boyer Commission on Educating Undergraduates in the Research University, 1998; National Research Council [NRC], 2003). Given that in 2009, only 21% of twelfth graders in the United States performed at or above the proficiency level in science (National Center for Education Statistics [NCES], 2009), the development of higher-level scientific-thinking skills in college poses a considerable challenge for both students and instructors. Repeated calls to reinvent science teaching and learning by the AAAS, the NRC, and the Boyer Commission on Educating Undergraduates in the Research University (AAAS, 1990, 1993, 2010; Boyer Commission, 1998; NRC, 2003) have been answered with many teaching innovations to promote student engagement and/or active learning strategies in science classes. These innovations include problem-based learning (Allen et al., 1996; Eberlein et al., 2008); process-oriented, guided-inquiry learning (Eberlein et al., 2008; Moog and Spencer, 2008); collaborative learning (Crouch and Mazur, 2001; Smith et al., 2009); peer-led team learning (Gosser and Roth, 1998; Stanger-Hall et al., 2010); a new emphasis on “scientific teaching” methodologies (Handelsman et al., 2004, 2007; Pfund et al., 2009); and the use of technological innovations, such as personal-response systems for student engagement and immediate in-class feedback (Caldwell, 2007; Smith et al., 2009), among others (Ebert-May and Brewer, 1997). Despite the increasing adoption of these innovative instruction methods over simple lecturing in college (66.3% of assistant professors and 49.6% of associate and full professors used student-centered and inquiry-based instruction methods in 2008; DeAngelo et al., 2009), and 99.6% of all university professors indicating that helping students develop critical-thinking skills is very important (DeAngelo et al., 2009), the outcomes for critical thinking have been disappointing so far. A recent study on student learning in U.S.
colleges and universities documented that 46% of college students did not gain critical-thinking skills during their first 2 yr of college, and 36% had not gained critical-thinking skills after 4 yr (Arun and Roksa, 2011). These data highlight the difficulties for both teaching and learning of critical-thinking skills in college, despite universal agreement on the importance of these skills. This raises the question whether this shortcoming is due to a lack of a critical-thinking challenge by instructors (Paul et al., 1997; Haas and Keeley 1998; Crowe et al., 2008; Zheng et al., 2008; Momsen et al., 2010; Arum and Roksa, 2011), student resistance to such a challenge (Keeley et al., 1995; Weimer, 2002; Arum and Roksa, 2011), or a combination of both. This study focuses on student resistance and possible influences on resistance, specifically the exam format used to assess student learning.

**Exam Format**

There has been considerable discussion on the advantages and disadvantages of different exam formats (Biggs, 1973; Simkin and Kuechler, 2005), both from a pedagogical (focused on learning outcomes) and from a practical (time and cost) perspective. In general, from an instructor standpoint, multiple-choice (MC) questions can be advantageous with respect to ease of scoring, perceived objectivity in grading, fast return of scores in large classes, and the capacity to ask more questions (Simkin and Kuechler, 2005). Limitations of MC exams include the work-intensive construction of high-quality question banks (Simkin and Kuechler, 2005), the difficulty of assessing critical-thinking skills (Martinez, 1999), and possibly false indication of student knowledge and understanding (Dufresne et al., 2002). While it is possible to write good critical-thinking (e.g., application, analysis, and evaluation) MC questions, they are usually difficult and time-intensive to create (Simkin and Kuechler, 2005), and synthesis (creative) skills cannot be assessed. In contrast, constructed-response (CR), such as fill-in-the-blank, short answer (SA), or essay questions that require students to create their own answer, can assess a wider range of thinking skills (Martinez, 1999), including critical thinking. In addition, CR questions give students the opportunity to express what they know (on all thinking levels). However, CR questions tend to be criticized for more subjective grading, intergrader variability, and time requirement for grading (Simkin and Kuechler, 2005). Students view MC exams as easier than essay exams and feel MC exams are easier to prepare for (require less time and effort), find the availability of options comforting, and like to be able to guess the right answer (Zeidner, 1987). Students also tend to expect knowledge questions from an MC exam and use surface (lower-level) learning when preparing for them (Scouller, 1998; Martinez, 1999). In contrast, students view essay exams as somewhat more appropriate for assessing the depth of their knowledge (Zeidner, 1987; Simkin and Kuechler, 2005). They may find grading more subjective, but some students believe that subjective grading could work to their advantage (Zeidner, 1987). Students also tend to expect more higher-level questions from CR exams and employ more deep-learning strategies in preparation for them (Scouller, 1998).

Finally, there is some inconclusive evidence for gender bias in assessment format. Some studies have found an advantage for males with MC questions, whereas others have found an advantage for males with CR questions (Bolger and Kellaghan, 1990; DeMars, 1998; Simkin and Kuechler, 2005).

**MC in Introductory Science Classes**

In a 2008 national survey, 33.1% of college instructors reported the use of MC exams (DeAngelo et al., 2009), but MC testing is especially common in introductory science classes at research universities, because of logistics and institutional policies (e.g., class size and grading support). Students in these classes tend to have been successful using memorization to prepare for and perform well on MC exams that emphasize lower-level thinking skills (e.g., Scholastic Assessment Test, advanced placement, other introductory science classes: Zheng et al., 2008). As a consequence, they have learned to associate MC questions with memorization (and other lower-level learning strategies: Scouller, 1998; Watters and Watters, 2007). This association likely undermines the credibility of the instructor and his or her attempts to convince introductory science students of the value of higher-level (critical) thinking, when students are tested with MC-only exams, even if these exams include higher-level MC questions. To test the hypothesis that the MC-only exam format hinders the development of higher-level thinking skills in introductory science students, I changed the exam format in a large introductory biology class with traditionally MC-only exams to include MC as well as several CR questions. I predicted that changing exam formats would result in 1) a change in how students studied and 2) a change in learning outcomes as assessed on the final exam, including 3) improved performance in critical (higher-level) thinking questions.

**Study Design**

The change in exam format was implemented in a large introductory biology class for biology majors (250–330 students per section). This section size is typical for general biology classes at research-intensive universities (Momsen et al., 2010). The class in this study was the second of two introductory biology classes in the major sequence, and it focused on organismal diversity, phylogenetics, the evolution of structures and functions in plants and animals, and ecology. Critical thinking was emphasized in this class, and Bloom’s taxonomy of thinking skills (Bloom, 1956; Anderson and Krathwohl, 2001) was taught to students as a communication and study tool during the first week of class. Critical thinking was defined in the framework of Bloom’s taxonomy (as Bloom levels 3–6: application, analysis, evaluation, and synthesis), because this hierarchical model of thinking skills creates specific expectations for practice and assessment at each thinking level (see also Simkin and Kuechler, 2005; Crowe et al., 2008). Furthermore, by using the term higher-level thinking, rather than the more abstract term critical thinking, it was the instructor’s intent to remind students that these are progressive thinking skills that build upon others and can be practiced.

After introducing Bloom’s taxonomy, the instructor explained to students that 25–30% of the questions on each exam would be asked at Bloom levels 3–5. As a result, students who desired to earn a grade of “C” or higher had to master these higher-level thinking skills (due to the MC exam format, level 6 [synthesis] was not assessed in the
the exams of the MC section consisted of 50 MC questions, and through interactive questioning during class. Throughout the semester students could practice their thinking skills by answering clicker questions, and subsequently discussing which evidence could be used to support or to eliminate different answer options (rather than being simply given the “correct” answer; see also Tanner and Allen, 2005). The students were also given question skeletons to design their own study questions at all Bloom levels, and instructed to practice the higher-level thinking skills during studying. However, despite these instructions (given to students every semester), students in this class usually struggle with the higher-level MC questions on the exams, and tend to perceive them as “tricky” (on the part of the instructor), rather than challenging (i.e., requiring higher-level thinking skills). After trying different approaches to convince the introductory biology students in this class of the value of practicing both lower- and higher-level thinking skills during studying (i.e., learning on all learning levels) while being assessed with a MC-only exam format (Stanger-Hall et al., 2011), I decided to test whether this exam format posed an obstacle for motivating students to practice higher-level thinking.

MATERIALS AND METHODS

During Spring 2009, two sections of this introductory biology class were offered, both taught by the same instructor (K.S.-H.). Instruction, assignments, and study tips were identical for both classes. The larger section (N = 282 consenting students) was assessed using the traditional MC-only exam format, and the smaller section (N = 192 consenting students) was assessed using a combination of MC, SA, and other CR questions (denoted by MC+SA hereafter). Students were not aware of the assessment format when signing up for one of the two class sections. Students in both sections answered four online surveys during the semester and took four exams and a comprehensive final exam. I used data from the four surveys (study behavior) and the cumulative final exam (learning outcomes) to test the predictions for this study.

Exam Format

The exams of the MC section consisted of 50 MC questions, the exams of the MC+SA section consisted of 30 MC and 3–4 SA questions. SA questions could usually be answered in three or four sentences or by labeling diagrams. For each exam, 25–30% of the questions were higher-level thinking questions (application or analysis), and for each exam, this proportion was the same for the two sections. Whenever possible, the MC questions in the MC-only section and the SA questions in the MC+SA section were designed to assess the equivalent content and thinking skills (see Supplemental Material 1 for examples). The students in this study were not specifically trained how to answer SA questions, and the instructor provided the same study recommendations and sample exam questions to both class sections. The sample exam questions (from previous semesters) were in MC format, but the instructions asked students to answer the questions as essay questions first (without looking at the answer options) and to use logical step-by-step reasoning to arrive at the answer. Only after reasoning out their answers were students to look at the answer options and chose the matching answer. In addition, students in both sections were assigned a two-page reading on how to answer an essay question. The handout used an example from class and gave the students different examples of student answers (different quality) that students were asked to compare and score (see Supplemental Material 2).

Scoring Rubrics for Exams. For each SA question the instructor worked with the grader (graduate teaching assistant) to create grading rubrics for content and reasoning skills (see Supplemental Material 3 for an example). For fill-in or labeling questions, only content was scored. The original rubrics were tested and modified with ~40 exam answers, and the resulting rubrics were used to grade all exams. During grading, the rubrics were further adjusted as necessary. The grader graded all exams to ensure grader consistency. The results of the MC questions were released within 1 d of the exam; the results of the SA questions were posted within 1 wk, along with the answer keys. Students did not receive individual feedback on their exam answers.

Regrade Requests. Students in both sections could submit a written regrade request for any exam question (explaining why they should receive credit and using scientific facts and reasoning) within 1 wk of the answer key being posted. These regrade requests were considered by the instructor and returned the following week with written feedback on the validity of the request.

Student Surveys

Students in both sections filled out four online surveys during the semester. They received 3 points of class credit (of 1100) for each survey. All surveys were given during non–exam weeks. The first survey (week 2) was given before exam 1, the second survey (week 6) was given before exam 2, the third survey (week 12) was given before exam 4, and the final survey was given during the last week (week 15) of class. During the first survey, students were asked which exam format they would prefer if given the choice (MC only, MC+SA, SA only), how many hours (per 7-d week) they usually studied for a science class during exam weeks and during non–exam weeks, and to report their current (start-of-the-semester) grade point average (GPA). In each of the four surveys, they were asked how much they had actually studied for the biology class during the previous 7-d week (a non–exam week), and which study behaviors they had used. The list of study behaviors contained nine cognitively passive (surface-learning) study behaviors and 13 cognitively active (deep-learning) study behaviors (Table 1), which had been developed from open-response student surveys in previous semesters and included study behaviors that had been recommended by the instructor at the beginning of the semester. Students could check as many study behaviors as they had used.

Although the term active learning is used in the literature (e.g., Prince, 2004) for both physically active (e.g., rewriting notes, making index cards) and cognitively active (e.g., making new connections, asking and answering new questions) learning behaviors, critical thinking cannot be achieved without the latter (Stanger-Hall et al., in preparation). For this
reason, the present analysis differentiated between cognitively passive (can be physically active) and cognitively active behaviors.

In a series of questions, the students were also asked to rate the strength of their own ability to perform certain assessment tasks (e.g., to remember facts and explanations, to apply what they have learned to different situations, to analyze problems, to evaluate different solutions to a problem) on a 5-point Likert scale (from 1: “I struggle with it” to 5: “I am excellent at it”). In addition, they were asked to rate the statement “I see the value of learning on all learning levels” on a Likert scale from 1 (completely disagree) to 5 (completely agree).

**Final Exams**

Student performance on the cumulative final exam was used to assess whether exam format affected student learning (final exams are not returned in this class, and none of the final exam questions had appeared on previous exams). For the MC section, the final exam consisted of 125 MC questions (90 of these questions were also asked in the MC+SA section) and three CR questions: a 24-item, fill-in question in table format; a 12-item, fill-in flowchart; and an extra credit SA question in short essay format. For the MC+SA section, the final exam consisted of 90 MC questions; a 24-item, fill-in question in table format; a 12-item, fill-in flowchart; and five SA questions (four of these were in short essay format). Students had 3 h to complete the final exam, and in both sections only very few (<10) students remained until the end of the allotted time.

For the final exam comparison between the two sections, I used the 90 identical MC questions (29 higher-level and 61 lower-level thinking questions; categorized based on class content and activities, assignments, and assigned reading); the cumulative 24-item, fill-in table (could be answered with lower-level skills, e.g., remembering from classes throughout the semester, but higher-level thinking would have helped retrieval and checking for errors); the 12-item, fill-in flowchart (remembering from class 2 wk before the final exam); and one common higher-level short essay question (extra credit for the MC class and part of the exam for the MC+SA class).

**GPA**

Due to class logistics, it was not possible to assign students randomly to the two exam formats, therefore previous student achievement (GPA) was used to account for potential student differences between treatment groups and was used as a covariate.

**Gender Differences**

To address potential gender differences in student performance on different exam questions, I coded the gender of the participating students based on their first names as male or female. Ambiguous names (e.g., Ashley, Kerry, Tyler, and names from other cultures) were not coded, and these students were not included in this analysis. The final gender sample size was N = 323 (N = 195 in MC class and N = 128 in MC+SA class).

**Student Evaluations**

Summary statistics for the anonymous end-of-semester class evaluations are reported. Evaluations were submitted by 207 students from the MC class and 130 students from the MC+SA class.

**Statistical Analysis**

I used SPSS version 19.0 (2011) for all quantitative statistical analyses. For each class, I tested all variables for normality (goodness of fit: Shapiro-Wilk test; SPSS 19.0). Only the total MC (90 questions) and the higher-level MC (29 questions) scores of the final exam were normally distributed. As a result, I report the results of nonparametric tests for all analyses. For the comparison of GPA, final exam scores, and study data between the two classes, I used the nonparametric Mann-Whitney U-test for independent samples. This is a test for both location and shape to test for differences between distributions of ranked variables. To compare data from the study surveys throughout the semester (repeated samples), I used the related-samples Wilcoxon signed-rank test. To correct for multiple comparisons (inflated type I error), I applied a false discovery rate correction (Benjamini and Hochberg, 1995) and report the adjusted p values. To compare student preferences for exam format and their attitudes regarding the value of learning on all learning levels between the two classes (i.e., exam formats), I applied a Pearson χ² test, using the data from the MC class to calculate the expected values for the MC+SA class. I used a Spearman correlation to assess a possible relationship between the change in value ratings (value

<table>
<thead>
<tr>
<th>Table 1.</th>
<th>List of cognitively passive and active learning behaviors that students reported in their study surveys</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitively passive learning behaviors</td>
<td>Cognitively active learning behaviors</td>
</tr>
<tr>
<td>I previewed the reading before class.</td>
<td>I asked myself: “How does it work?” and “Why does it work this way?”</td>
</tr>
<tr>
<td>I came to class.</td>
<td>I drew my own flowcharts or diagrams.</td>
</tr>
<tr>
<td>I read the assigned text.</td>
<td>I broke down complex processes step-by-step.</td>
</tr>
<tr>
<td>I rewrote my class notes.</td>
<td>I wrote my own study questions.</td>
</tr>
<tr>
<td>I made index cards.</td>
<td>I reorganized the class information.</td>
</tr>
<tr>
<td>I highlighted the text.</td>
<td>I compared and contrasted.</td>
</tr>
<tr>
<td>I looked up information.</td>
<td>I fit all the facts into a bigger picture.</td>
</tr>
<tr>
<td>I asked a classmate or tutor to explain the material to me.</td>
<td>I tried to figure out the answer before looking it up.</td>
</tr>
<tr>
<td>I reviewed my class notes.</td>
<td>I closed my notes and tested how much I remembered.</td>
</tr>
<tr>
<td>I wrote my name on the exam.</td>
<td>I asked myself: “How are individual steps connected?” and “Why are they connected?”</td>
</tr>
<tr>
<td>I used a calculator.</td>
<td>I drew and labeled diagrams from memory and figured out missing pieces.</td>
</tr>
<tr>
<td>I reviewed my class notes.</td>
<td>I asked myself: “How does this impact my life?” and “What does it tell me about my body?”</td>
</tr>
<tr>
<td>I used Bloom’s taxonomy to write my own study questions.</td>
<td>I highlighted the text.</td>
</tr>
<tr>
<td>I rewrote my notes.</td>
<td>I fit all the facts into a bigger picture.</td>
</tr>
<tr>
<td>I reorganized the class information.</td>
<td>I compared and contrasted.</td>
</tr>
<tr>
<td>I asked myself: “How are individual steps connected?” and “Why are they connected?”</td>
<td>I drew and labeled diagrams from memory and figured out missing pieces.</td>
</tr>
<tr>
<td>I asked myself: “How does this impact my life?” and “What does it tell me about my body?”</td>
<td>I highlighted the text.</td>
</tr>
</tbody>
</table>

Vol. 11, Highlights of 2012
rating after four exams minus value rating before the first exam) and the change in higher-level ability ratings (average in higher-level ability ratings after four exams minus average in higher-level ability ratings before the first exam). For a comprehensive analysis of the influences of previous student achievement (GPA), exam format, and gender on final exam performance, I conducted a two-way analysis of covariance (ANCOVA) with GPA as a covariate. For all variables we report means and standard error of the mean (mean ± SEM). All reported statistical results are based on two-tailed tests and significance levels of \( p < 0.05 \).

**Sample Sizes for Analysis.** In the MC class, 282 students consented to participate in this study, 231 students reported their start-of-semester GPA, and 195 students were identified (based on their first names) as male \((N = 78)\) or female \((N = 117)\). A total of 242 students finished the class (took the final exam). Of these, 172 students took all four study surveys (for longitudinal comparison) and reported on study times and study activities. Thus, the reported results are based on sample sizes of 242 (final exam performance), 231 (GPA), 195 (gender), or 172 (study time and activities). When asked their preferred exam format (MC only, MC+SA, or SA only) at the beginning of the semester, 60% of the students in this class preferred MC only, 36% preferred MC+SA, and 4% preferred SA only.

In the MC+SA class, 192 students consented to participate in this study, 155 students reported their start-of-semester GPA, and 128 students were identified (based on their first names) as male \((N = 63)\) or female \((N = 65)\). A total of 164 students finished the class (took the final exam). Of these, 121 students took all four study surveys and reported on study times and study activities. Thus, the reported results are based on sample sizes of 164 (final exam performance), 155 (GPA), 128 (gender), or 121 (study time and activities). When asked their preferred exam format (MC only, MC+SA, or SA only) at the beginning of the semester, 58% of the students in this class preferred MC only, 36% preferred MC+SA, and 6% preferred SA only. This was not significantly different from the MC class \(\chi^2 = 2.117, p = 0.347\).

**ANCOVA for GPA, Gender, and Exam Format.** To control for the influence of previous student achievement and gender on final exam performance I included GPA as a covariate in a two-way (gender and exam format) ANCOVA.

**RESULTS**

**Exam Format and Studying**

**Study Time for Science Classes.** The students in the two sections did not differ in their reported study times for a science class in general (exam weeks: Mann-Whitney \(U = 9850.5, p = 0.430\); non-exam weeks: Mann-Whitney \(U = 9350.5, p = 0.213\)), and there was no difference between male \((N = 125)\) and female \((N = 172)\) students (exam weeks: Mann-Whitney \(U = 11,528.5, p = 0.280\); non-exam weeks: Mann-Whitney \(U = 11,961, p = 0.092\)). All students combined \((N = 323)\) reported an average study time of 3.16 ± 0.098 h/wk during non-exam weeks (range: 0 min to more than 9 h), and an average of 8.39 ± 0.211 h/wk for exam weeks (range: 2 h to more than 20 h) for a science class.

**Study Time for Biology Class.** In line with their other science classes, students in the MC section reported an average study time of \((mean ± SEM) 3.31 ± 0.18 h/wk for the biology class in this study, and students in the MC+SA section reported on average of 3.31 ± 0.21 h/wk (Figure 1A) for the second week of class. After the first exam, students in both sections increased their study time (Table 2), but neither of these changes was significant (Table 3). Between the second and the fourth exam, students increased their weekly study time significantly to an average of 4.33 ± 0.23 h in the MC section and an average of 4.60 ± 0.24 h in the MC+SA section. Before the final exam, both sections significantly decreased their weekly study time (Table 3). There was no significant difference in study time between the two sections at any point in the semester (Figure 1A and Table 3).

**Study Behavior.** At the beginning of the semester, students in the MC section reported an average of 3.93 ± 0.84 cognitively passive learning behaviors (of nine options), while the students in the MC+SA section reported an average of 3.91 ± 0.17 cognitively passive learning behaviors during studying for this class (non-exam week; Table 2); this was not significantly different (Mann-Whitney \(U = 10,249, p = 0.824\), Table 2). This trend continued for the remainder of the semester (Figure 1B): students in the two sections did not change their cognitively passive learning behaviors...
Table 2. Reported study time and study activities through the semester\(^a\)

<table>
<thead>
<tr>
<th>Change over time</th>
<th>Time 1</th>
<th>Time 2</th>
<th>Time 3</th>
<th>Time 4</th>
<th>Compare 1–2</th>
<th>Compare 2–3</th>
<th>Compare 3–4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Mean ± SEM</td>
<td>Mean ± SEM</td>
<td>Mean ± SEM</td>
<td>Mean ± SEM</td>
<td>Z</td>
<td>p</td>
</tr>
<tr>
<td>MC</td>
<td>172</td>
<td>3.31 ± 0.18</td>
<td>3.59 ± 0.19</td>
<td>4.33 ± 0.23</td>
<td>3.74 ± 0.22</td>
<td>4.374</td>
<td>&gt;0.05</td>
</tr>
<tr>
<td>Passive</td>
<td>3.93 ± 0.13</td>
<td>3.66 ± 0.27</td>
<td>3.77 ± 0.14</td>
<td>3.58 ± 0.15</td>
<td>4.861</td>
<td>&gt;0.05</td>
<td>3.699</td>
</tr>
<tr>
<td>Active</td>
<td>1.91 ± 0.17</td>
<td>2.31 ± 0.18</td>
<td>3.20 ± 0.23</td>
<td>2.55 ± 0.22</td>
<td>3.915.5</td>
<td>&lt;0.05*</td>
<td>5.966</td>
</tr>
<tr>
<td>MC+SA</td>
<td>121</td>
<td>3.31 ± 0.22</td>
<td>3.33 ± 0.20</td>
<td>4.60 ± 0.24</td>
<td>3.27 ± 0.22</td>
<td>1.840</td>
<td>&gt;0.05</td>
</tr>
<tr>
<td>Passive</td>
<td>3.91 ± 0.17</td>
<td>3.62 ± 0.17</td>
<td>3.82 ± 0.16</td>
<td>3.85 ± 0.17</td>
<td>1.600</td>
<td>&gt;0.05</td>
<td>2.644</td>
</tr>
<tr>
<td>Active</td>
<td>2.54 ± 0.21</td>
<td>3.17 ± 0.28</td>
<td>3.87 ± 0.28</td>
<td>3.42 ± 0.29</td>
<td>3.159</td>
<td>&lt;0.05*</td>
<td>3.043.5</td>
</tr>
</tbody>
</table>

*Statistically significant at \(p < 0.05\).

\(^a\)Study time is reported in hours per week (mean ± SEM), study activities (passive and active) are reported as number of activities used in that week (mean ± SEM). Time 1: before exam 1; time 2: before exam 2; time 3: before exam 4; time 4: before final exam. Sample sizes (MC: 172; MC + SA: 121) are based on students that answered all four surveys. Pairwise comparisons: related samples Wilcoxon signed-rank test (\(Z\) values). \(p\) values are reported after Benjamini-Hochberg correction for multiple comparisons.

Table 3. Section differences in study time and study activities through the semester\(^a\)

<table>
<thead>
<tr>
<th>Study behavior comparison</th>
<th>Before exam 1</th>
<th>Before exam 2</th>
<th>Before exam 4</th>
<th>Before final exam</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mann-Whitney U</td>
<td>p</td>
<td>Mann-Whitney U</td>
<td>p</td>
</tr>
<tr>
<td>MC (N)</td>
<td>172</td>
<td>121</td>
<td>10,381</td>
<td>0.973</td>
</tr>
<tr>
<td>MC+SA (N)</td>
<td>172</td>
<td>121</td>
<td>10,249</td>
<td>0.824</td>
</tr>
<tr>
<td>Passive</td>
<td>172</td>
<td>121</td>
<td>11,264.5</td>
<td>&lt;0.05*</td>
</tr>
</tbody>
</table>

*Significant at the \(p < 0.05\) level.

\(^a\)An independent sample Mann-Whitney \(U\)-test was used. \(p\) values are reported after Benjamini-Hochberg correction for multiple comparisons.
significantly (Table 2), and they did not differ significantly from one another in the number of cognitively passive learning behaviors they reported at any point in the semester (Table 3). In contrast, the reported cognitively active learning behaviors (of 13 options) increased significantly from the beginning of the semester (MC: 1.91 ± 0.172; MC+SA: 2.54 ± 0.212) to the second exam (MC: 2.31 ± 0.177; MC+SA: 3.17 ± 0.276) and from the second to the fourth exam (MC: 3.20 ± 0.226; MC+SA: 3.87 ± 0.283) in both sections (Figure 1B). Between the fourth and the final exam, there was a significant drop in cognitively active learning behaviors in both sections (Table 3). These changes in cognitively active learning behaviors remained significant (at \( p < 0.05 \)) or marginally significant (at \( p = 0.05 \)) after correcting for multiple comparisons (\( N = 3 \)). Even though both sections changed their cognitively active learning behavior throughout the semester, students in the MC+SA section reported significantly more cognitively active learning behaviors than the students in the MC section in three of the four surveys (Figure 1B). All differences between the two sections remained significant (at the \( p < 0.05 \) level) after correcting for multiple (\( N = 4 \)) comparisons.

**Exam Format and Final Exam Performance**

**Total Final Exam Scores.** Within each section (MC and MC+SA), students did not differ significantly in how they performed on the final exam regardless of the preference for exam format (MC only, MC+SA, SA only) they had stated at the beginning of the semester (independent samples Kruskal-Wallis test \( df = 2 \) for MC section: \( 2.265, p = 0.322 \); for MC+SA section: \( 12.06, p = 0.547 \)). But the students in the MC+SA section scored significantly higher (67.34%) on the final exam than the students in the MC-only section (63.82%, Mann-Whitney \( U = 23,622, p = 0.001 \)).

**CR Question Scores.** The MC and MC+SA sections had three CR questions in common on the final exam. The students in the MC+SA section scored significantly higher (67.27 ± 1.00) on these three CR questions (SA, fill-in table, and fill-in flowchart) than the students in the MC section (61.97 ± 0.85, Mann-Whitney \( U = 24,540.5, p < 0.001 \); Figure 2A and Tables 4 and 5).

**MC Question Scores.** Students in the MC and MC+SA sections answered 90 identical MC questions on their final exams. The students in the MC+SA section scored significantly higher (67.35%) on these 90 MC questions than the students in the MC section (64.23%, Mann-Whitney \( U = 23,095.5, p = 0.005 \); Figure 2A). This difference was mostly due to a significantly better performance on the higher-level MC questions: students in the MC+SA section scored significantly higher (64.4%) on the higher-level questions than the students in the MC section (59.54%; Mann-Whitney \( U = 24,035, p < 0.001 \)). The difference between the two sections on lower level MC questions (68.76% vs. 66.46%) was marginally significant (Mann-Whitney \( U = 22,114, p = 0.05 \)). All differences remained significant after adjustment for multiple comparisons (\( N = 3 \); Table 5).

**Previous Student Achievement (GPA).** At the beginning of the semester students in the MC section reported an average GPA of 3.3 ± 0.025, and students in the MC+SA section reported an average GPA of 3.2 ± 0.035 (Table 4). This was not significantly different (Mann-Whitney \( U = 15,869, p = 0.053 \), Table 5).

**Gender.** Overall, male students scored significantly higher (67.10%) than female students (64.18%, Mann-Whitney \( U = 1104.5, p = 0.032 \)) on the final exam (Table 4). Male students did not significantly differ from female students in the CR questions on the final exam (63.95% vs. 64.34%, Mann-Whitney \( U = 12,942.5, p = 0.893 \)), but they performed significantly better on both lower-level (69.19% vs. 66.05%, Mann-Whitney \( U = 10,989, p = 0.027 \)) and higher-level (64.88% vs. 60.14%, Mann-Whitney \( U = 10,206.5, p = 0.002 \)) MC questions (significant after adjustment for multiple comparisons: \( N = 3 \); Figure 2B). With reference to the overall student scores (mean ± SD: lower level = 67.39 ± 11.58; higher level = 61.5 ± 13.2), male students performed 0.273 SD units better on the lower level, and 0.359 SD units better on the higher-level MC questions than female students. These gender differences were at least partially influenced by differences in exam format. When considering the two class sections separately (Table 6), the gender differences in average scores were less pronounced (and not significant: Table 7), but in each section,
Table 4. Summary statistics of GPA and final exam performance

<table>
<thead>
<tr>
<th>Exam format and gender</th>
<th>GPA Mean ± SEM</th>
<th>FE total (%)</th>
<th>CR total (%)</th>
<th>MC total (%)</th>
<th>MC lower (%)</th>
<th>MC higher (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC (242)</td>
<td>3.30 ± 0.025</td>
<td>63.82 ± 0.68</td>
<td>61.97 ± 0.85</td>
<td>64.23 ± 0.71</td>
<td>66.46 ± 0.74</td>
<td>59.54 ± 0.82</td>
</tr>
<tr>
<td>MC+SA (164)</td>
<td>3.20 ± 0.035</td>
<td>67.34 ± 0.13</td>
<td>67.27 ± 1.00</td>
<td>67.35 ± 0.87</td>
<td>68.76 ± 0.89</td>
<td>64.40 ± 1.05</td>
</tr>
<tr>
<td>Male (141)</td>
<td>3.25 ± 0.039</td>
<td>67.10 ± 0.39</td>
<td>63.95 ± 1.13</td>
<td>67.80 ± 0.93</td>
<td>69.19 ± 0.95</td>
<td>64.88 ± 1.14</td>
</tr>
<tr>
<td>Female (182)</td>
<td>3.29 ± 0.029</td>
<td>64.18 ± 0.78</td>
<td>64.34 ± 0.97</td>
<td>64.14 ± 0.82</td>
<td>66.05 ± 0.86</td>
<td>60.14 ± 0.93</td>
</tr>
</tbody>
</table>

GPA (mean ± SEM) is based on smaller sample sizes (MC: 231; MC + SA: 155; male: 129; female: 176) of students who reported their start-of-semester GPA. Final exam performance (mean ± SEM) is reported as percent of total for the different performance categories: FE total (final exam score), CR total (3 CR questions), MC total (all 90 MC questions), MC lower (61 lower-level MC questions), MC higher (29 higher-level MC questions).

Table 5. Exam format and gender differences in GPA and final exam performance

<table>
<thead>
<tr>
<th>Exam format</th>
<th>GPA Mean ± SEM</th>
<th>FE total (%)</th>
<th>CR total (%)</th>
<th>MC total (%)</th>
<th>MC lower (%)</th>
<th>MC higher (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC by gender</td>
<td>3.16 ± 0.034</td>
<td>62.56 ± 1.0</td>
<td>62.18 ± 1.22</td>
<td>62.64 ± 1.04</td>
<td>1.036 ± 0.21</td>
<td></td>
</tr>
<tr>
<td>MC + SA by gender</td>
<td>3.21 ± 0.053</td>
<td>64.18 ± 0.78</td>
<td>68.23 ± 1.5</td>
<td>66.83 ± 1.22</td>
<td>1.003 ± 0.25</td>
<td></td>
</tr>
</tbody>
</table>

*Significant at the *p* < 0.05 level. (*) Marginally significant at *p* = 0.05.
GPA (mean ± SEM) is based on smaller sample sizes (MC: 231; MC + SA: 155; male: 129; female: 176) of students who reported their start-of-semester GPA. An independent sample Mann-Whitney U-test was used (Mann-Whitney *U* decimals were rounded up from 0.5 to 1.0). *p values are reported after Benjamini-Hochberg correction for multiple comparisons.

Table 6. Summary statistics of GPA and final exam performance by exam format and gender

<table>
<thead>
<tr>
<th>Exam format by gender</th>
<th>GPA Mean ± SEM</th>
<th>FE total (%)</th>
<th>CR total (%)</th>
<th>MC total (%)</th>
<th>MC/CR ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC male (78)</td>
<td>3.31 ± 0.017</td>
<td>65.05 ± 1.16</td>
<td>60.54 ± 1.54</td>
<td>66.05 ± 1.16</td>
<td>1.129 ± 0.02</td>
</tr>
<tr>
<td>MC female (117)</td>
<td>3.34 ± 0.034</td>
<td>62.56 ± 1.0</td>
<td>62.18 ± 1.22</td>
<td>62.64 ± 1.04</td>
<td>1.036 ± 0.21</td>
</tr>
<tr>
<td>MC + SA male (63)</td>
<td>3.16 ± 0.066</td>
<td>69.63 ± 1.40</td>
<td>68.17 ± 1.51</td>
<td>69.94 ± 1.4</td>
<td>1.037 ± 0.17</td>
</tr>
<tr>
<td>MC + SA female (65)</td>
<td>3.32 ± 0.053</td>
<td>64.18 ± 0.78</td>
<td>68.23 ± 1.5</td>
<td>66.83 ± 1.22</td>
<td>1.003 ± 0.25</td>
</tr>
</tbody>
</table>

GPA (mean ± SEM) is based on smaller sample sizes of students who reported their start-of-semester GPA. Final exam performance (mean ± SEM) is reported as percent of total for the different performance categories: FE total (final exam score), CR total (3 CR questions), MC total (all 90 MC questions), MC/CR ratio (MC%/CR% = 1.0 if students do equally well on MC and CR).

Table 7. Differences in GPA and final exam performance by exam format and gender

<table>
<thead>
<tr>
<th>Exam format by gender</th>
<th>GPA Mann-Whitney U</th>
<th><em>p</em></th>
<th>FE total Mann-Whitney U</th>
<th><em>p</em></th>
<th>CR total Mann-Whitney U</th>
<th><em>p</em></th>
<th>MC total Mann-Whitney U</th>
<th><em>p</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>MC by gender</td>
<td>4,001</td>
<td>0.751</td>
<td>4,030</td>
<td>0.167</td>
<td>48,203</td>
<td>0.437</td>
<td>3,325</td>
<td>0.001*</td>
</tr>
<tr>
<td>MC + SA by gender</td>
<td>1,867</td>
<td>0.635</td>
<td>1,750</td>
<td>0.156</td>
<td>2,025</td>
<td>0.914</td>
<td>1,558</td>
<td>0.020*</td>
</tr>
</tbody>
</table>

*Significant at the *p* < 0.05 level.
An independent sample Mann-Whitney U-test was used (Mann-Whitney U decimals were rounded up from 0.5 to 1.0).
male students tended to perform better on the MC questions than the female students in the same section, but this was not the case for the CR questions.

**Exam Format, GPA, Gender, and Final Exam Performance.** GPA significantly influenced student performance on the final exam (ANCOVA $F = 85.0$, $p < 0.001$), explaining an estimated 22.1% of the total variance. After accounting for GPA (evaluated at 3.28), both gender ($F = 8.27$, $p = 0.004$) and exam format ($F = 29.33$, $p < 0.001$) significantly influenced final exam scores, with gender explaining 2.7% of the total variance, and exam format explaining 8.9%. The estimated marginal means of the final exam scores (after accounting for GPA) were 63.38 ± 0.704 for students in the MC-only class, compared with 69.43 ± 0.860 for students in the MC + SA class (and 68.0 ± 0.83 for male and 64.81 ± 0.731 for female students; Figure 2).

**Exam Format and Student Attitudes toward Higher-Level Thinking.** At the beginning of the semester (students knew their exam format but had not taken any exams yet), students in the MC + SA section tended to “see the value of learning on all learning levels” significantly more than the students in the MC section ($\chi^2 = 12.131$, $p(2) = 0.0023$). In the MC section, 57.3% of students agreed or strongly agreed that they saw the value of learning on all learning levels (Figure 3), and 11.1% of students disagreed or strongly disagreed with the statement. In the MC + SA section, 72.1% of students agreed or strongly agreed that they saw the value of learning on all learning levels, and 7.4% of students disagreed or strongly disagreed. Students in both sections changed their rating significantly between the beginning and the end of the semester (MC: Wilcoxon $Z = 1385.5$, $p < 0.001$; MC + SA: Wilcoxon $Z = 646$, $p < 0.001$). At the end of the semester, students in both sections responded very similarly ($\chi^2 = 0.196$, $p(2) = 0.907$), with an overall lower value rating for learning on all learning levels. To test the hypothesis that this may be a response to being challenged—and struggling—with critical-thinking questions in the exams of this class, I analyzed whether this change in value was related to a change in how students rated their own ability in the higher-level (critical-thinking) assessment tasks after getting feedback on exams. Across both classes, students who rated their ability in performing these higher-level tasks more highly at the end of the semester (after taking four exams) compared with the beginning (before taking any exam), also tended to rate the value of learning on all learning levels more highly than they did at the beginning of the semester. Similarly, students who rated their own ability in the higher-level (critical-thinking) tasks lower after taking four exams, also tended to rate the value of learning on all learning levels lower at the end of the semester. This resulted in a significant positive correlation between student perceptions of their own critical-thinking ability and how they valued learning on all learning levels (Spearman’s rho = 0.263, $p < 0.001$, $N = 335$).

**Exam Format and Student Evaluations.** Even though students in the MC + SA section learned significantly more than students in the MC section, they did not like being assessed with CR questions. In the anonymous end-of-semester class evaluations, the students in the MC + SA section rated the fairness of grading in the course much lower than did the students in the MC-only section (Figure 4). In their written comments, the MC + SA students attributed their low ratings to the fact that they had to answer SA questions, while their friends in the other section did not (the different grading scales in the two sections did not make a difference to them). Students in both sections commented on the emphasis on higher-level thinking in the introductory biology class in this study. For example, some students noted that the instructor should “just teach biology” rather than emphasize higher-level thinking skills.
DISCUSSION

The MC-Only Exam Format Poses an Obstacle for Critical Thinking

The purpose of this study was to assess whether an MC-only exam format might hinder the development of higher-level (critical) thinking skills in introductory science students. The answer is a convincing yes. The MC-only exam format seemed to undermine the instructor’s efforts to convince students of the importance of critical-thinking skills, even though 25–30% of the MC questions assessed higher-level thinking. Simply knowing that they would be assessed with SA questions in addition to MC questions, significantly more students in the MC + SA section (72% vs. 57%) reported that they saw the value of learning on all learning levels at the beginning of the semester (before taking any exam). This perception was associated with a different approach to studying and a significantly better performance on the final exam. This illustrates the powerful role of perceptions of assessment in the learning process (Scouller, 1998; Watters and Watters, 2007). It is well known that students have different expectations for MC and CR exams, and as a result study differently in preparation for these exams (e.g., Scouller, 1998). However, in most previous studies that compared actual performance on MC versus CR exams, MC questions were treated as a homogeneous entity without consideration of the question level (see also Simkin and Kuechler, 2005; Kuechler and Simkin, 2010), and as a result, these studies may have compared performance on different cognitive levels rather than performance due solely to the format of the questions. In contrast, in the present study, each exam (both formats) was designed to include 25–30% higher-level thinking questions, and the students were made aware of that before taking any exams.

Interestingly, students in the MC + SA section did not study more than the students in the MC-only section. Students in both sections spent on average considerably less time studying (3 h per non-exam week) than was recommended by the instructor (2 h per hour class time or 6 h/wk). This is in line with national data: college students spend on average a total of 15 h/wk on studying, or about 7% of their time in a 5-d week (Arum and Roksa, 2011). These data also reflect national trends of declining study times in college students (Babcock and Marks, 2011): full-time college students in 1961 allocated on average 24.4 h/wk to studying, while in 2003 students spent on average 14.4 h/wk (10 h fewer).

Instead of studying more, the students in the MC + SA section used their study time more effectively for practicing higher-level thinking. Students in both sections reported a similar number of cognitively passive (surface) learning behaviors (~3.5) during studying (Figure 1), and the average number of reported cognitively active (deep) learning behaviors increased in both sections in response to their exams. This shows that students will respond with more active learning if challenged, even in the MC-only format. However, the students in the MC + SA section consistently reported more cognitively active learning behaviors in non-exam weeks (Figure 1) than the students in the MC-only section, and this difference in study behavior translated into significantly better student performance on the cumulative final exam. The somewhat puzzling decrease in study time before the cumulative final exam (Figure 1A) could be explained if most students (incorrectly and against instructions) assumed that the final exam would mostly consist of repeated questions from previous exams and were planning on memorizing the old exam questions during the week of the final exam, and/or if an extraordinary amount of semester papers and/or lab reports from other classes was due during that week, and students waited until the last minute to work on these at the cost of their final exam preparation. The significant drop in self-reported active learning behaviors during the last class week (Figure 1B) supports a shift to memorization.

The MC + SA students significantly outperformed the MC students on all final exam measures (MC and CR items, Tables 4 and 5). It could be argued that the MC students did not take the SA question too seriously, because it was an extra credit question; however, the MC + SA students significantly outperformed the MC students in the other question types (fill-in table, fill-in flowchart, and MC) as well. Most importantly, the significantly better performance of the MC + SA students on the MC questions was mostly due to a significantly better performance on the higher-level (critical-thinking) MC questions. This further supports the hypothesis that the MC-only exam format indeed discourages the practice of critical-thinking skills in introductory science classes, while the addition of CR questions encourages it.

While the change to a mixed exam format in introductory science classes requires a commitment by colleges and universities to provide adequate grading support, this investment would be a cost-effective strategy to significantly improve the critical-thinking skills of college students.

Exam Format and Student Evaluations

The clear student preference for assessment with MC questions (and student perception of MC questions being easier to answer and thus less effort to prepare for) is reflected in the assessment literature (Simkin and Kuechler, 2005). Due to the mixed exam format, mistakes in reasoning were more obvious for the students in the MC + SA section and likely contributed to their less-favorable student evaluations of both the class and the instructor at the end of the semester (Kearney and Plax, 1992; Keeley et al., 1995). But even though many students in the MC + SA section disliked the experience, they learned significantly more, including critical-thinking skills, than the students in the MC-only section. This illustrates the limited use of student evaluations as a measure of actual student learning, and suggests that student ratings should not be overinterpreted, especially if students are asked to practice new thinking skills (McKeachy, 1997).

Overcoming Resistance

Student resistance to learning seems to be a common occurrence in college classrooms (Burroughs et al., 1989; Kearney and Plax, 1992). For example, the comments of students in both sections that the instructor should “just teach biology,” rather than emphasize thinking skills, seems to be a typical expression of student resistance to a critical-thinking challenge (e.g., Keeley et al., 1995). This resistance (defined by Keeley et al., 1995 as any student behavior that hinders their development into critical thinkers) was also expressed in students spending on average 50% less time studying than was recommended by the instructor, insistence on using mainly
cognitively passive study strategies, and downgrading the value of learning on all learning levels when struggling on exams.

As pointed out by Karpicke and coworkers (Karpicke et al., 2009), some students may be under the illusion of competence and believe that they know the material better than they actually do when they rely purely on their subjective learning experience (e.g., their fluency of processing information during rereading and other passive study strategies). As students adjusted their own competency ratings with feedback from exams, students who downgraded their higher-level thinking skills tended to like the idea of learning on all levels less, and students who reported an increase in their higher-level thinking skills tended to value learning on all learning levels more.

Given the increased learning gains with the mixed exam format, an important question for both instructors and students is how to overcome student misgivings (e.g., Kearney and Plax, 1992) about the learning process to further maximize learning gains. In the present study, possible sources for student resistance included: 1) different exam formats in different sections, 2) expectation to practice unfamiliar thinking skills, and 3) overestimation of own critical-thinking ability. To reduce these influences, ideally all introductory science classes should implement a mixed exam format. This would not only improve student learning, but would also reduce student resistance associated with the perception of unfairness in grading due to different exam formats. In addition, all college classes should emphasize higher-level (critical) thinking skills (AAAS, 1990, 1993, 2010; Boyer Commission, 1998; NRC, 2003). This would greatly reduce student resistance to critical thinking in individual college classes. However, this is presently not the case (Crowe et al., 2008; Arum and Roksa, 2011), possibly due to lingering faculty resistance toward teaching critical thinking (Haas and Keeley, 1998), and unfamiliarity of faculty with how to teach critical-thinking skills (DeAngelo et al., 2009). Finally, the resistance component due to discomfort associated with facing one’s own limitations (e.g., when failing to reason out an answer on an exam) could be reduced if students were trained to construct written answers with proper reasoning. In the current study, exam format accounted for 9% of the variance on the final exam performance. This occurred without practice opportunities for constructing arguments and reasoning out answers in a written format (e.g., through graded homework assignments). By adding such opportunities (requiring additional teaching assistant support) the critical-thinking gains would be expected to be even higher, while student resistance to critical thinking should be reduced. With more practice opportunities and individual feedback, students should gain competency faster, and more students should end the semester with a (realistically) higher rating of their critical-thinking skills and a more positive attitude toward higher-level learning. Ideally, combining these approaches would refocus student energies away from resisting toward practicing their higher-level thinking.

Gender Bias
A potentially troubling issue for any instructor is the possibility that exam format per se could create a performance bias beyond student achievement. In the present study, male students performed significantly better on the MC questions of the final exam than female students. An important question is whether this is an accurate measure of student achievement or whether this is due to a bias in assessment format. Research has shown that society-specific gender stereotypes predict sex differences in science performance (Nosek, 2009), and these differences in the approach to science are hard to change due to student (and school) focus on grades over engagement with the material (Carlone, 2004). In 2008, even though male and female U.S. twelfth-graders did not differ significantly in their science scores, male scores tended to be higher than female scores (NCES, 2009), and more male students (26%) scored above the proficiency level than female students (19%). Male students also tended to have completed more science courses (biology, chemistry, and physics) in high school (NCES, 2009), which has been shown to be a good predictor for science success in college (Muller et al., 2001; Arum and Roksa, 2011).

As a consequence, at least some of the gender difference in the MC questions on the final exam in this study seems to be based on differences in achievement. However, if entirely due to achievement differences, the MC differences should also be reflected in the other question formats. In the present study, male students tended to perform better than female students on both assessment formats on the final exam, but they performed relatively better on the MC questions than on the CR questions, resulting in significant gender differences for MC questions only. This suggests that there may be at least some inherent bias toward male students in the MC question format (e.g., through differences in “testwiseness” [Zimmerman and Williams, 2003] and/or male students being more willing to guess than female students [Ben-Shakhar and Sinai, 2005]). Whatever the reason, the change from MC-only exam formats in introductory science classes to mixed exam formats would not only increase student learning and higher-level thinking in general, but would also remove a potential handicap for female students in introductory science classes and possibly encourage their pursuit of a career in the STEM disciplines.

ACKNOWLEDGMENTS
This work was supported by a University of Georgia System Board of Regents STEM grant, which funded the hiring and training of a graduate student teaching assistant for grading the exams in the MC+SA section. This study was conducted under the guidelines of IRB # 2007-10197. I thank Tom Koballa for advice in the planning phase of this study, Julie Palmer for inspiring the essay exercise, and the UGA Science Education Research Group for feedback on an earlier version of the manuscript. This is a publication of the UGA Science Education Research Group.

REFERENCES


Graduate teaching assistants (GTAs) are used extensively as instructors in higher education, yet their status and authority as teachers may be uncertain to undergraduates, to administrators, and even to the GTAs themselves. This study explored undergraduate perception of classroom instruction by GTAs and professors to identify factors unique to each type of instructor versus the type of classes they teach. Data collection was via an online survey composed of subscales from two validated instruments, as well as one open-ended question asking students to compare the same class taught by a professor versus a GTA. Quantitative and qualitative results indicated that some student instructional perceptions are specific to instructor type, and not class type. For example, regardless of type of class, professors are perceived as being confident, in control, organized, experienced, knowledgeable, distant, formal, strict, hard, boring, and respected. Conversely, GTAs are perceived as uncertain, hesitant, nervous, relaxed, laid-back, engaging, interactive, relatable, understanding, and able to personalize teaching. Overall, undergraduates seem to perceive professors as having more knowledge and authority over the curriculum, but enjoy the instructional style of GTAs. The results of this study will be used to make recommendations for GTA professional development programs.

INTRODUCTION

There is a growing dependence on contingent instructors (part-time, non-tenure-track faculty and graduate teaching assistants [GTAs]) at research universities (Johnson, 2011); specifically, Baldwin and Wawrzynski (2011) and Jaeger (2008) indicate that these contingent instructors may contribute to roughly half of the instructional staff. According to one survey in the biological sciences discipline in the United States (n = 65), GTAs are responsible for teaching 71% of undergraduate laboratory sections at their comprehensive institutions and 91% at their research institutions (Sundberg et al., 2005). Similarly, Rushin et al. (1997) found that 97% of 153 graduate schools surveyed in the United States used GTAs to teach laboratories and/or lectures in biology courses. While there may be variability in these numbers from institution to institution, they highlight the reliance of higher education on GTA employment.

The majority of contingent instructors provide instruction to lower-division courses, making the likelihood of contact with them greatest for first-year students. Given the importance of first-year coursework to student retention, understanding the impact of contingent instructors on student learning is critical (Benjamin, 2002; Jaeger, 2008). However, few studies have been conducted on the quality of teaching students receive from contingent instructors, and those that have been conducted report conflicting results or are limited by sample size (Umbach, 2007).

For instance, Bolge (1995) reported that contingent instructors do not differentially affect student outcomes, because there were no differences in final grades. Yet Johnson (2011) determined that contingent instructors typically give higher grades. It has also been reported that full-time, tenure-track faculty devote proportionately more time to students than do contingent instructors (Benjamin, 2002). Similarly, Jaeger (2008) noted that contingent instructors are generally less accessible and less available, even though students note that out-of-class interactions are most important for their education. Umbach (2007) suggested that contingent instructors...
are typically less effective in how they work with undergraduates than tenured/tenure-track faculty. Yet Johnson (2011) reported that instructor type does not impact student retention, while O’Neal et al. (2007) found that GTA enthusiasm increased the likelihood of student retention. Baldwin and Wawrzynski (2011) posited that there is sufficient evidence to merit concern about the teaching quality of contingent instructors, and they are proponents of targeted professional development strategies for different types of contingent instructors. For this reason, this study focuses on identifying factors that undergraduates perceive as different between GTAs and professors in order to make suggestions specifically for GTA professional development.

GTAs as Contingent Instructors

While GTAs are used extensively as undergraduate instructors, they are admitted to universities to pursue graduate education at their institutions. Hence, some define a GTA as a graduate student pursuing a master’s or doctoral degree who is used part-time to provide instruction to undergraduates, while others have gone as far as calling GTAs “donkeys of the department” (Marincovich et al., 1998), due to their immense workload, level of responsibility, and restricted autonomy (Park and Ramos, 2002). Thus, GTAs occupy a unique position in the academic system as both researchers and faculty/professors in training, whether by choice or not (Golde, 1998; Marincovich et al., 1998; Park, 2002; Park and Ramos, 2002; Muzaka, 2009). The perceived role of a GTA at an institution varies based on who is asked. For instance, undergraduates perceive GTAs as holding a status between students and academics, while GTAs see themselves as students with teaching responsibilities (Park, 2002; Muzaka, 2009). Faculty members consider GTAs to be research students who are also academic apprentices (Park, 2002; Muzaka, 2009). Their role is sometimes puzzling to administrators and policy makers, who are often unsure whether to classify GTAs as students or staff (Flora, 2007). Some studies attempting to clarify GTA responsibilities have gone as far as calling GTAs “donkeys of the department,” due to their immense workload, level of responsibility, and restricted autonomy (Park and Ramos, 2002). Thus, GTAs occupy a unique position in the academic system as both researchers and faculty/professors in training, whether by choice or not (Golde, 1998; Marincovich et al., 1998; Park, 2002; Park and Ramos, 2002; Muzaka, 2009).

GTA Professional Development

Given the reliance on GTAs for a majority of science laboratory teaching at U.S. universities, the need for effective professional development is a necessity. GTAs want preparation and guidance throughout their teaching experiences to improve not only their instructional ability but also their overall teaching experience (Bond-Robinson and Rodrigues, 2006). Given that GTAs may be future faculty members at academic institutions, there is also a need to carefully mentor GTAs and provide them with advanced instructional assignments as they progress in their teaching experience (Braxton et al., 1995); however, GTA assignments are typically made to cover departmental needs and not the professional development needs of future faculty members (Austin, 2002). The reality is that GTAs often feel unprepared for their teaching assignments (Dudley, 2009), and most universities offer no formal professional training at all (Rushin et al., 1997).

A study by Rushin et al. (1997) surveyed graduate schools (n = 153) in the United States to determine what training opportunities they offer to GTAs. Although the most common response was that no formal training was required, the second most common approach was a pre-academic-year workshop (Rushin et al., 1997). The next most common GTA development opportunities, in decreasing prevalence, were: a semester-long college teaching seminar, a formal college teaching course, and training by a professor (Rushin et al., 1997). Typical professional development activities included teaching multiple courses during the graduate program, videotaping for self-evaluation, written training manuals, weekly meetings prior to teaching, and semester-long courses (Rushin et al., 1997), activities similar to those suggested by graduate students to improve their professional development as teachers (Nyquist et al., 1999). These findings are also similar to Marincovich et al. (1998), who found that the topics most commonly addressed in GTA professional development included teaching in a nontraditional setting, ethical issues, communication skills, developing reflection habits (including evaluating teaching skills), and obtaining frequent feedback (including close mentorship).

The Golde and Dore (2001) Pew Charitable Trusts study also asked doctoral students (n = 4114) from 11 Arts and Sciences departments about preparation for teaching. More than half of these doctoral students indicated they were required to teach during their degree program, and these students also reported that they were interested in, prepared for, and confident in teaching laboratory sections and lecture courses and in leading discussions (Golde and Dore, 2001). Yet Golde and Dore (2001) were unable to determine whether this confidence was merited, since graduate students clearly noted that their programs did not prepare them for these instructional roles.

Several researchers have indicated that even if departments provide instructional training, GTAs often receive minimal amounts of the pedagogical and content instruction information that they need (Marincovich et al., 1998; Shannon et al., 1998; Luft et al., 2004). Frequently absent from their training is the necessary background on delivering specific curricula, course planning, and assessment, or skills such as interdisciplinary connections, interactive pedagogy, instructional design, and teamwork (Marincovich et al., 1998; Shannon et al., 1998; Luft et al., 2004).

The offering of GTA professional development indicates that universities understand the importance of training; however, the diversity of training available and the reported lack of consistency in GTA training manuals (Lowman and Mathie, 1993) also indicates that more research needs to be done on what aspects of training are most beneficial to GTAs. Shannon et al. (1998), for instance, found that techniques such as microteaching, coteaching, and practice simulations were the most important aspects of developing teaching effectiveness in GTAs. They argue, however, that more progress will be made in determining the aspects essential to training when agreement is reached regarding the purpose and definition of GTA professional development (Shannon et al., 1998). This study responds to this call by identifying factors that undergraduates perceive to be different between GTAs and
characteristics that may impact the effectiveness of in-
struction. This study resulted in the generalizations that GTA
ratings are not impacted by undergraduate student major and
gender of the GTA, yet they are influenced by degree held (higher degrees result in higher ratings), teaching ex-
perience, and GTA age (those in their late twenties are rated
higher than those in their early twenties; Bos et al., 1980).

Project Rationale

Previous studies have found that there are instructional char-
acteristics of GTAs that impact the perception of teaching and
learning at universities; however, there have been few direct
comparisons of student perceptions of professors and GTAs,
particularly those employing the same methodologies on the
same undergraduate student population. The goal of this
study is to provide a quantitative and qualitative comparison
of professors and GTAs from the perspective of undergrad-
uate students at one institution in order to identify factors
useful for GTA professional development. Although profes-
sors and GTAs may teach similar content and students, they
often teach different types of classes, which could confound
the identification of variations specific to each type of instruc-
tor. Thus, this study is designed to identify factors unique to
each type of instructor, versus the type of classes they teach,
by collecting data on undergraduate student perception of
each type of instructor in each of two classroom situations
(laboratory and discussion). The hypothesis for this study is
that undergraduates will perceive a difference between pro-
fessors and GTAs related to the instructor and not just the
classroom environment.

MATERIALS AND METHODS

Data Collection

Data were collected through an online survey (hosted by
 surveymonkey.com) composed of subscales from two pub-
lished and validated instruments, the College and University
Classroom Environment Inventory (CUCEI) and the Ques-
tnaire of Teacher Interaction (QTI). These instruments were
chosen for this study as they best captured the previously identified instructional characteristics of GTAs (e.g., uncertainty, approachable, etc.), and both
had been used individually and together in previous studies in
the college environment (Treagust and Fraser, 1986; Coll et al., 2002). The CUCEI was designed to incorporate the di-
mensions identified by Moos (1979) in his work on classroom
environment and includes the subscales of: Personalization,
Innovation, Student Cohesiveness, Satisfaction, Task Orienta-
tion, Innovation, and Individualization (Treagust and Fraser,
1986; Coll et al., 2002). The QTI was originally designed in
the Netherlands to explore how individuals mutually influence
each other and consists of 77 items; however, it was shortened to 48 items for use in the Australian science educa-
tion environment (Coll et al., 2002). The shortened QTI covers
the subscales of: Leadership, Understanding, Uncertain,
Admonishing, Helpful/Friendly, Student Responsibility and
Freedom, Dissatisfied, and Strict (Coll et al., 2002).

Using both the CUCEI and QTI instruments in their entirety
would result in a survey requiring responses to 97 items, not

Vol. 11, Highlights of 2012

Does Instructor Type Matter?
all of which were necessarily relevant to the study. Given the voluntary nature of the survey, and since not all of the subscales were directly related to potential differences between GTAs and professors (based on previous research), it was determined that only certain subscales of each instrument would be used in the study. Therefore, this project utilized only the CUCEI subscales of Personalization, Involvement, Task Orientation, and Individualization. For the QTI, only the subscales of Leadership, Uncertain, Helpful/Friendly, and Strict were used. The subscales were chosen to best capture aspects that were similar to the results of previous research, such as positive aspects of GTA instructors (e.g., approachability, informality, enthusiasm, and less intimidating, which related to the CUCEI and QTI subscales of Helpful/Friendly, Personalization, Task Orientation, Leadership, and Strict), and negative aspects (e.g., lack of experience, control, and content knowledge, and limited contact time, which related to the CUCEI and QTI subscales of Leadership, Task Orientation, Uncertain, Individualization, and Involvement; Bos et al., 1980; Park, 2002; Dudley, 2009; Muzaka, 2009). The subscale questions, as well as the Likert choices undergraduates were given for each, are shown in Tables 1 and 2.

The survey was sent via email in September 2010 (4 weeks into the Fall semester) to undergraduate students enrolled in majors and nonmajors general biology courses (introductory biology, plant biology, biodiversity, cell biology, genetics, and ecology) at a large research institution in the southern United States. Each of the selected courses has a lecture and laboratory component; the laboratories are all taught by GTAs, who teach two or three laboratory sections per semester, and the lecture classes are all taught by PhD-level faculty (tenure-line or lecturers). Undergraduate students received the email, which contained a description of the project and a link to the survey, from their lecture instructor on behalf of the researcher. The communication was done via the lecture and not the laboratory, because more undergraduates would receive the information at one time and much of the course communication is routinely done through the lecture instructor. The project description indicated that responding to the survey was voluntary and was not related to the course grade;
additionally, undergraduates were assured that their instructor would not see their responses to the survey. Throughout the survey, undergraduates were asked to focus on generic and not specific instructors.

Prior to beginning the survey, undergraduate students responded to demographic questions about their gender, enrollment status, major, native language, and other biology courses they had completed. Student participants then answered the same set of survey questions but were assigned (by their last name for each lecture class) to only one of four scenarios: 1) imagine a small discussion class (20–25 students) taught by a professor, 2) imagine a small discussion class (20–25 students) taught by a GTA, 3) imagine a lab class (20–25 students) taught by a professor, or 4) imagine a lab class (20–25 students) taught by a GTA.

After respondents were informed of their scenario, they responded to items from the CUCEI. The original directions for the CUCEI were included, and students were told to respond to each item using a four-point scale of “strongly disagree,” “disagree,” “agree,” and “strongly agree.” The order the items were presented in was similar to the methodology in Coll et al. (2002) and Treagust and Fraser (1986) to be consistent with previous studies. After students completed the CUCEI items, they were presented with directions for the QTI. The response options were on a five-point scale ranging from “0” (the instructor never displays this behavior) to “4” (the instructor always displays this behavior). The QTI items were presented in an order similar to that used by Coll et al. (2002). Students were unable to change their responses once they had completed a page. Undergraduates were also asked to confirm their scenario at the end of the survey to allow the researcher to remove responses from undergraduates who might have forgotten their scenario midsurvey. After responding to the CUCEI and QTI items, all undergraduate students then responded to the same open-ended question: If the same class was taught by a professor versus a graduate teaching assistant, how do you think the classes would be different?

No incentives were offered for participation, and all procedures were reviewed and approved by the Institutional Review Board for Human Subjects.

Survey Validity and Reliability

Because this project used two pre-existing instruments (both previously validated) in the same survey, with four different scenarios, internal consistency/reliability estimates were calculated for each subscale, as well as for each subscale for each scenario, using Cronbach’s alpha. For the overall data, these values ranged from 0.655 to 0.876, depending on the subscale, and were higher than those reported for previous uses of the instruments in college environments, except for the subscale of Individualization (Coll et al., 2002; Treagust and Fraser, 1986). For the subscales by scenarios, the reliability measures varied from 0.580 to 0.911 for each subscale. From these results, it was judged that the instruments were reliable to use for the project.

Validity of the open-response question was judged via face validity; in that most students responded to the question and gave responses that were consistent with the intent of the question. No students wrote that they did not understand what the question was asking.

Data Analysis

Quantitative Data. Data from the CUCEI and QTI subscales were coded independently, due to differences in the Likert scaling. The CUCEI responses were coded in the following manner: “strongly agree” = 5, “agree” = 4, “disagree” = 2, and “strongly disagree” = 1. Nonpositive items (see Table 1) were “flipped” and then coded: “strongly agree” = 1, “agree” = 2, “disagree” = 4, or “strongly disagree” = 5. The QTI was coded in the following manner: responses of “always” were scored as “4,” and then coding progressed down to “never,” which was scored as “0.” No questions from the QTI were nonpositive. Student responses for each scenario were then compiled for comparison.

The quantitative data were analyzed using nonparametric methodology, because the Likert-type choices are ordinal. In ordinal data, there is no guarantee that students perceive the difference between intervals on the point scale as equal distances (e.g., “agree” and “strongly agree” are at the same distance as “agree” and “disagree”). This theoretical lack of equal distances violates assumptions for parametric methodology, requiring the data to be analyzed via nonparametric tests (Huck, 2008).

For each of the CUCEI and QTI items, student responses among each of the four scenarios were compared using Kruskal-Wallis one-way analysis of variance (ANOVA) tests (SPSS Statistics 19.0). Items in which statistically significant differences were found ($\alpha < 0.05$) were then compared using pairwise comparisons (Mann-Whitney U-tests, with Bonferroni approach to control for type I error across tests $[\alpha < 0.013]$). Pairwise comparisons were performed to determine whether the differences were due to the classroom setting or instructor type; the former were considered to be “classroom variables” and the latter were considered to be “instructor variables.” Significant classroom variables were identified by comparing professors in a discussion class with professors in a lab class and GTAs in a discussion class to GTAs in a lab class, while instructor variables were identified by comparing professors in a discussion class with GTAs in a discussion class and professors in a lab class to GTAs in a lab class.

The majority of undergraduate participants were first-year students (53%); consequently, Kruskal-Wallis one-way ANOVA tests were performed to determine whether student enrollment status affected the results. It was found that responses from first-year students did not differ from second-, third-, fourth-year, and beyond students; therefore, undergraduate students were grouped together for all analyses.

Qualitative Data. Open-ended responses ($n = 127$) to the question about differences between a class taught by a GTA versus a professor were first sorted into responses that indicated there would be no difference and those that stated there would be a difference. Undergraduate student responses in which differences were perceived ($n = 110$) then underwent thematic analysis using a “grounded theory” approach (Corbin and Strauss, 1990; LeCompte, 2000) in which the researchers let the results emerge from the data without preconceived ideas about what students might articulate. Responses were analyzed independently by two researchers (each author of this paper) who read and reread the responses and took notes on the differences undergraduates articulated about the two different types of instructors. Factors that arose consistently were grouped and given a name (key word) and
description, and then each researcher tallied the number of times those key words or descriptions appeared in student responses. Each researcher then compiled her findings, and only then did the researchers compare their results. The identified key words were then grouped into themes via discussion between the two researchers until both were in agreement.

As an example of this process, student responses yielded the key word “relatable” as one possible emergent characteristic of GTAs. Student responses that indicated that GTAs identified with them, related to them, knew what it was like to be in a student’s shoes, were classified into this category. This key word of “relatable” was then merged with other key words (such as “respect,” “boring,” and “approachable”) into a theme that was entitled “relationship,” because they all were thought to be articulating how students and instructors interact on a personal level (one on one, not just in a classroom delivery setting).

Reliability of these results was first obtained by the concordance of key words and descriptions between the two authors. Even if the key words were different, the ideas captured from the student responses were the same, and the researchers discussed the final title of the key word to reach consensus. The results also aligned with several of the quantitative results of this study (e.g., uncertain and nervous), as well as the results of previous studies (e.g., approachable, relatable, uncertain, nervous, and limited control [Park, 2002; Dudley, 2009; Muzaka, 2009]). These multiple sources of verification of the results were considered evidence of the reliability of the findings.

RESULTS

Participants

The survey was sent to a potential undergraduate pool of 2586 undergraduate students. From this pool, 387 began the survey (15.0%), while 225 undergraduates completed the survey (8.7%). Undergraduates who were minors or who could not remember their scenario at the end of the survey were removed, leaving 184 total respondents (7.1%). This response rate was likely a result of the survey being voluntary, with no incentives for participation. Of the 184 respondents, survey completion for each scenario was: 59 for a discussion class taught by a professor; 43 for a discussion class taught by a GTA; 38 for a lab class taught by a professor; and 44 for a lab class taught by a GTA. There were 167 undergraduates who responded to the open-ended question. After removing responses in which the instructor they were referring to was indeterminable (e.g., “They are more understanding”), 127 respondents remained for analysis.

Individuals who completed the survey were mostly freshman (first year; 53%), non-biology majors (74%), female (72%), and native English speakers (95%). Second- and third-year students comprised 21 and 18%, respectively, of the respondents, with 8% more being fourth year or beyond. Twenty-six percent of the students were biology majors, with 4% concentrating in ecology and evolutionary biology, 10% in biochemistry and cellular and molecular biology, and 3% in microbiology. Most of the respondents were currently enrolled in a majors’ biology course (62%), while the rest were currently enrolled in nonmajors courses. The majority of respondents had not completed another biology course (63%); however, 37% had completed at least one other semester-long lecture/lab biology course, such as first semester nonmajors biology (11%), biodiversity (17%), and cell biology (13%). A complete summary of the demographics for overall survey respondents, as well as demographics for each scenario, is shown in Table 3.

Quantitative Analysis

Item Analysis. Items from each survey in which significant differences in the medians among the scenarios were found (Kruskal-Wallis; \( \alpha < 0.05 \)) are shown in Table 4. Descriptive statistics for each of these items for each scenario are shown in Table 5. These items included: students knowing what to do in class \( (\chi^2 = 7.95, df = 3, n = 183, p = 0.047) \), having a say in how class time is spent \( (\chi^2 = 7.821, df = 3, n = 184, p = 0.050) \), class being disorganized \( (\chi^2 = 8.264, df = 3, n = 184, p = 0.041) \), students being allowed to choose activities and how they will work \( (\chi^2 = 9.004, df = 3, n = 181, p = 0.029) \), students having opportunities to express their opinions in class \( (\chi^2 = 10.976, df = 3, n = 184, p = 0.012) \), activities being clearly and carefully planned \( (\chi^2 = 13.09, df = 3, n = 184, p = 0.004) \), the teacher talking enthusiastically about the subject \( (\chi^2 = 24.098, df = 3, n = 184, p = 0.000) \), the teacher being uncertain \( (\chi^2 = 16.20, df = 3, n = 184, p = 0.001) \), the teacher being hesitant \( (\chi^2 = 10.628, df = 3, n = 183, p = 0.014) \), and the teacher knowing what to do \( (\chi^2 = 17.319, df = 3, n = 184, p = 0.001) \). These significant differences spanned items from five of the eight subscales of the two instruments, including Task Orientation, Individualization, Involvement, Leadership, and Uncertain.

Pairwise Comparisons. To determine whether the significant differences among scenarios were a result of classroom or instructor variables, pairwise comparisons were completed between classroom scenarios and then between instructor scenarios for each significant item. The comparisons among classroom scenarios (a GTA teaching a lab vs. teaching a discussion, or a professor teaching a lab vs. teaching a discussion) found no differences between any of the scenarios (Table 4). These results indicate that undergraduates in this study perceive instructor qualities to be the same in discussion classrooms as in lab classrooms.

For instructor variables (a GTA vs. a professor teaching the same type of class), there were no significant differences in responses between a professor teaching a discussion class and a GTA teaching a discussion class. However, undergraduates perceived differences in a professor teaching a lab class and a GTA teaching a lab class. In the situation of the professor and GTA both teaching a lab class, significant differences in undergraduate responses occurred in the factors of planning \( (U = 570.500, n = 81, p = 0.004) \), enthusiasm \( (U = 373.500, n = 81, p = 0.000) \), uncertainty \( (U = 553.500, n = 81, p = 0.005) \), hesitation \( (U = 518.000, n = 81, p = 0.002) \), and knowing what to do \( (U = 464.500, n = 81, p = 0.000) \).
more enthusiastic and have activities more clearly and carefully planned, while GTAs are more uncertain, hesitant, and act as if they do not know what to do. These results are similar to those obtained from the qualitative data (see below), except the factor of enthusiasm, which was not mentioned by undergraduates in the open-ended responses.

### Qualitative Analysis

While ~13% \((n = 17)\) of the undergraduate student respondents indicated that they would not perceive a difference in a class taught by a professor versus a GTA, analysis of the responses from undergraduates who did perceive differences \((n = 110)\) generated two overall themes: factors relating to the teaching realm (further subdivided into themes of delivery technique and classroom atmosphere) and factors related to the personal realm (particularly with regard to relationship). The delivery technique, classroom atmosphere, and relationship themes are explained in Table 6 by using the key words and main descriptions researchers used to characterize undergraduate perceptions of the differences between GTAs and professors in the open-ended question. The numbers in parentheses after each bold descriptive key word indicate the number of times responses were grouped into those key words. A single student response could contain anywhere from zero to three key words (mean = 1.4 key words). These perceived differences between GTAs and professors were identified as instructor variables and not classroom variables, since participants were asked about how GTAs and professors would differ in teaching the same type of class.

The theme of delivery technique includes student responses that seemed to reflect the characteristics of the instructors as teachers, specifically how they deliver the course material to students and how confident the instructor appears in the classroom. This theme incorporated responses that referenced classroom control, organization and preparedness, knowledge level, and teaching experience. Undergraduates described professors as being more experienced, structured, confident, knowledgeable, organized, and in control in the classroom as compared with GTAs, who undergraduates described as more hesitant, nervous, and uncertain. For instance, student 58 stated, “TAs tend to be unorganized, without a strong curriculum to back them up, nor do they have the teaching experience that gives them the courage to stand

---

### Table 3. Demographic percentages for overall study participants, as well as by scenario*

<table>
<thead>
<tr>
<th>What is your gender?</th>
<th>Overall</th>
<th>Discussion professor</th>
<th>Discussion GTA</th>
<th>Lab professor</th>
<th>Lab GTA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>28%</td>
<td>29%</td>
<td>30%</td>
<td>29%</td>
<td>23%</td>
</tr>
<tr>
<td>Female</td>
<td>72%</td>
<td>71%</td>
<td>70%</td>
<td>71%</td>
<td>77%</td>
</tr>
<tr>
<td>What is your current enrollment status?</td>
<td>Overall</td>
<td>Discussion professor</td>
<td>Discussion GTA</td>
<td>Lab professor</td>
<td>Lab GTA</td>
</tr>
<tr>
<td>First year</td>
<td>53%</td>
<td>58%</td>
<td>49%</td>
<td>39%</td>
<td>61%</td>
</tr>
<tr>
<td>Second year</td>
<td>21%</td>
<td>22%</td>
<td>28%</td>
<td>26%</td>
<td>9%</td>
</tr>
<tr>
<td>Third year</td>
<td>18%</td>
<td>15%</td>
<td>14%</td>
<td>21%</td>
<td>23%</td>
</tr>
<tr>
<td>Fourth year</td>
<td>6%</td>
<td>3%</td>
<td>5%</td>
<td>11%</td>
<td>7%</td>
</tr>
<tr>
<td>Fifth year and/or beyond</td>
<td>2%</td>
<td>2%</td>
<td>2%</td>
<td>3%</td>
<td>0%</td>
</tr>
<tr>
<td>What is your major?</td>
<td>Overall</td>
<td>Discussion professor</td>
<td>Discussion GTA</td>
<td>Lab professor</td>
<td>Lab GTA</td>
</tr>
<tr>
<td>Biology</td>
<td>9%</td>
<td>7%</td>
<td>12%</td>
<td>3%</td>
<td>14%</td>
</tr>
<tr>
<td>Ecology and evolutionary biology</td>
<td>4%</td>
<td>2%</td>
<td>2%</td>
<td>8%</td>
<td>5%</td>
</tr>
<tr>
<td>Biochemistry, cellular, and molecular biology</td>
<td>10%</td>
<td>10%</td>
<td>5%</td>
<td>13%</td>
<td>11%</td>
</tr>
<tr>
<td>Microbiology</td>
<td>3%</td>
<td>0%</td>
<td>2%</td>
<td>3%</td>
<td>7%</td>
</tr>
<tr>
<td>Other</td>
<td>74%</td>
<td>81%</td>
<td>79%</td>
<td>74%</td>
<td>64%</td>
</tr>
<tr>
<td>Is English your native language?</td>
<td>Yes</td>
<td>95%</td>
<td>93%</td>
<td>98%</td>
<td>100%</td>
</tr>
<tr>
<td>No</td>
<td>4%</td>
<td>7%</td>
<td>2%</td>
<td>0%</td>
<td>2%</td>
</tr>
<tr>
<td>What other core courses have you completed?</td>
<td>Overall</td>
<td>Discussion professor</td>
<td>Discussion GTA</td>
<td>Lab professor</td>
<td>Lab GTA</td>
</tr>
<tr>
<td>First semester nonmajors biology</td>
<td>11%</td>
<td>15%</td>
<td>5%</td>
<td>13%</td>
<td>11%</td>
</tr>
<tr>
<td>Second semester nonmajors biology</td>
<td>8%</td>
<td>8%</td>
<td>7%</td>
<td>11%</td>
<td>7%</td>
</tr>
<tr>
<td>First semester plant biology</td>
<td>2%</td>
<td>2%</td>
<td>2%</td>
<td>5%</td>
<td>0%</td>
</tr>
<tr>
<td>Second semester plant biology</td>
<td>2%</td>
<td>2%</td>
<td>2%</td>
<td>3%</td>
<td>0%</td>
</tr>
<tr>
<td>Biodiversity</td>
<td>17%</td>
<td>14%</td>
<td>14%</td>
<td>24%</td>
<td>23%</td>
</tr>
<tr>
<td>Honors biodiversity</td>
<td>1%</td>
<td>3%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Cell biology</td>
<td>13%</td>
<td>7%</td>
<td>12%</td>
<td>18%</td>
<td>18%</td>
</tr>
<tr>
<td>General genetics</td>
<td>6%</td>
<td>3%</td>
<td>7%</td>
<td>8%</td>
<td>7%</td>
</tr>
<tr>
<td>General ecology</td>
<td>1%</td>
<td>0%</td>
<td>0%</td>
<td>3%</td>
<td>2%</td>
</tr>
<tr>
<td>No others</td>
<td>63%</td>
<td>66%</td>
<td>69%</td>
<td>50%</td>
<td>66%</td>
</tr>
<tr>
<td>Current course enrollment</td>
<td>Overall</td>
<td>Discussion professor</td>
<td>Discussion GTA</td>
<td>Lab professor</td>
<td>Lab GTA</td>
</tr>
<tr>
<td>First semester nonmajors biology</td>
<td>38%</td>
<td>49%</td>
<td>38%</td>
<td>29%</td>
<td>30%</td>
</tr>
<tr>
<td>First semester plant biology</td>
<td>4%</td>
<td>2%</td>
<td>10%</td>
<td>3%</td>
<td>2%</td>
</tr>
<tr>
<td>Biodiversity</td>
<td>34%</td>
<td>31%</td>
<td>38%</td>
<td>29%</td>
<td>41%</td>
</tr>
<tr>
<td>Honors biodiversity</td>
<td>3%</td>
<td>5%</td>
<td>0%</td>
<td>3%</td>
<td>5%</td>
</tr>
<tr>
<td>Cell biology</td>
<td>8%</td>
<td>7%</td>
<td>5%</td>
<td>18%</td>
<td>5%</td>
</tr>
<tr>
<td>General genetics</td>
<td>7%</td>
<td>5%</td>
<td>7%</td>
<td>3%</td>
<td>11%</td>
</tr>
<tr>
<td>General ecology</td>
<td>7%</td>
<td>2%</td>
<td>5%</td>
<td>16%</td>
<td>7%</td>
</tr>
</tbody>
</table>

*Overall \(n = 184\); Discussion professor \(n = 59\), Discussion GTA \(n = 43\); Lab professor \(n = 38\); Lab GTA \(n = 44\).
Table 4. Significant results for the Kruskal-Wallis ($\alpha < 0.05$) and Mann-Whitney $U$ ($\alpha < 0.013$ with the Bonferroni correction) nonparametric tests

<table>
<thead>
<tr>
<th>Item</th>
<th>Kruskal-Wallis one-way ANOVA</th>
<th>Mann-Whitney $U$, $p$ value</th>
<th>Classroom variables</th>
<th>Instructor variables</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\chi^2$</td>
<td>$p$ value</td>
<td>DP vs. LP</td>
<td>DG vs. LG</td>
</tr>
<tr>
<td>Students are allowed to choose activities and how they will work. (CUCEI)</td>
<td>9.004</td>
<td>0.029</td>
<td>0.076</td>
<td>0.067</td>
</tr>
<tr>
<td>Students know exactly what has to be done in our class. (CUCEI)</td>
<td>7.95</td>
<td>0.047</td>
<td>0.920</td>
<td>0.663</td>
</tr>
<tr>
<td>Students have a say in how class time is spent. (CUCEI)</td>
<td>7.821</td>
<td>0.050</td>
<td>0.100</td>
<td>0.077</td>
</tr>
<tr>
<td>This is a disorganized class. (CUCEI)</td>
<td>8.264</td>
<td>0.041</td>
<td>0.717</td>
<td>0.301</td>
</tr>
<tr>
<td>There are opportunities for students to express their opinions in this class. (CUCEI)</td>
<td>10.976</td>
<td>0.012</td>
<td>0.166</td>
<td>0.039*</td>
</tr>
<tr>
<td>Activities in this class are clearly and carefully planned. (CUCEI)</td>
<td>13.09</td>
<td>0.004</td>
<td>0.696</td>
<td>0.344</td>
</tr>
<tr>
<td>This teacher is hesitant. (QTI)</td>
<td>10.628</td>
<td>0.014</td>
<td>0.091</td>
<td>0.679</td>
</tr>
<tr>
<td>This teacher talks enthusiastically about her/his subject. (QTI)</td>
<td>24.098</td>
<td>0.000</td>
<td>0.118</td>
<td>0.017*</td>
</tr>
<tr>
<td>This teacher seems uncertain. (QTI)</td>
<td>16.206</td>
<td>0.001</td>
<td>0.359</td>
<td>0.166</td>
</tr>
<tr>
<td>This teacher acts as if she/he does not know what to do. (QTI)</td>
<td>17.319</td>
<td>0.001</td>
<td>0.103</td>
<td>0.554</td>
</tr>
</tbody>
</table>

*Significant items from the CUCEI (6) and QTI (4) shown in the order they were presented to students. Significant instructor variables are bolded in the bottom right-hand corner. Letters in parentheses indicate who had the higher values for each item. $D =$ Discussion, $L =$ Lab, $P =$ Professor, $G =$ GTA. *Mann-Whitney $U$, $p$ value significant before Bonferroni correction ($\alpha < 0.05$); **Mann-Whitney $U$, $p$ value significant with Bonferroni correction ($\alpha < 0.013$).
<table>
<thead>
<tr>
<th></th>
<th>Professor Discussion</th>
<th>GTA Discussion</th>
<th>Professor Lab</th>
<th>GTA Lab</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean ± standard deviation</td>
<td>Med</td>
<td>Max</td>
<td>Min</td>
</tr>
<tr>
<td>Students are allowed to choose activities and how they will work.</td>
<td>2.88 ± 1.10</td>
<td>2</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Students know exactly what has to be done in our class.</td>
<td>3.19 ± 1.03</td>
<td>4</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Students have a say in how class time is spent.</td>
<td>3.03 ± 1.17</td>
<td>4</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>This is a disorganized class.</td>
<td>4.12 ± 1.02</td>
<td>4</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>There are opportunities for students to express opinions in this class.</td>
<td>4.00 ± 1.07</td>
<td>4</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Activities in this class are clearly and carefully planned.</td>
<td>3.90 ± 0.84</td>
<td>4</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>This teacher is hesitant.</td>
<td>1.02 ± 0.99</td>
<td>1</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>This teacher talks enthusiastically about her/his subject.</td>
<td>3.27 ± 0.83</td>
<td>3</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>This teacher seems uncertain.</td>
<td>0.81 ± 0.94</td>
<td>1</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>This teacher acts as if she/he does not know what to do.</td>
<td>0.61 ± 1.02</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 6. Summary of themes (delivery technique, classroom atmosphere, and relationship) obtained from qualitative data in both the teaching and personal realms for GTAs and professors.

<table>
<thead>
<tr>
<th>Realm</th>
<th>Theme</th>
<th>GTA</th>
<th>Professor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teaching</td>
<td>Delivery technique</td>
<td>Hesitant, nervous, uncertain (11), and unsure how to begin teaching</td>
<td>Organized and structured (17), confident (10), in control (10), prepared for questions, with previous teaching experience (15), and greater knowledge (21)</td>
</tr>
<tr>
<td>Classroom atmosphere</td>
<td>Relaxed and laid-back (9), interactive, engaging (5), personalized, and having open student-instructor interactions (3)</td>
<td>Distant and formal (9), strict (13), serious, harder (2), with higher expectations and standards</td>
<td></td>
</tr>
<tr>
<td>Personal</td>
<td>Relationship</td>
<td>Comfortable approaching GTAs (8) and that GTAs are relatable (19) and understanding (5)</td>
<td>Intimidating and boring (2), and out of touch (6), yet undergraduates respect (14) professors</td>
</tr>
</tbody>
</table>

Numbers in parentheses are how often the bolded key words emerged from the data set.

before students and confidently teach the material.” Meanwhile, student 153 elaborated on this by saying, “Professors are usually more confident and know the subject or lab better. GTAs are always a little uncertain and don’t always answer questions well.” The key words that comprised this theme were the most common responses by undergraduates to the open-ended question, especially in terms of the knowledge (n = 21), structure/organization (n = 17), and experience (n = 15) of professors being greater than that of GTAs.

The classroom atmosphere theme grouped student comments that seemed to be about student-instructor relationships within the context of classroom instruction. Undergraduates in the study articulated that the classroom atmosphere with a professor was more distant, strict, and formal, as compared with that of a GTA, whose classroom was seen as more relaxed, laid-back, and personalized. Furthermore, some undergraduates expressed that professor-led classrooms had higher expectations and were more serious than GTA-led classrooms, which some undergraduates indicated were more interactive and engaging. For instance, student 60 said, “The professor might be more strict rather than a graduate teaching assistant who might be a little more laid back because they can relate to the stress of college and how they might have had a bad teacher or hard professor.” Student 156 explained, “Professors are typically old and out of touch with the students. They are set in their ways about making them look younger and less serious.” In terms of the relationship aspect, undergraduates noted their respect for professors (n = 14) and their ability to relate to GTAs (n = 19) as the primary descriptors.

**DISCUSSION**

Bos et al. (1980) suggested caution when comparing different studies regarding GTA instructional abilities with ones for ranked faculty members; our study has addressed this concern through comparison of GTAs and professors in the same study using the same sample population and instruments. This study has further addressed the call for developing targeted professional development (Shannon et al., 1998; Baldwin and Wawrzynski, 2011) by identifying aspects that distinguish GTAs from faculty and that could be the focus of professional development.

We hypothesized that undergraduates would perceive differences between GTAs and professors that were independent of classroom variables and found evidence that supported this hypothesis. Undergraduates in this study perceived professors as being more structured, confident, in control, organized, experienced, knowledgeable, distant, formal, strict, serious, hard, boring, out of touch, and respected than GTAs. Conversely, GTAs were perceived as more uncertain, hesitant, nervous, relaxed, laid-back, engaging, interactive, relatable, understanding, and able to personalize teaching than professors.

Significant differences between instructor type were found for five of the 52 items surveyed in the quantitative portion of this study, while no significant classroom variables were identified. Thus, although students do see many similarities between professors and GTAs, there are also core factors that undergraduates at one institution say are different between GTAs and professors independent of what classes they teach. These results were supported by the qualitative data analysis, which independently confirmed four of the five instructor differences from the quantitative analysis. In addition, the open-response format allowed students to add additional variables that were not identified from the survey.

Although it remains to be seen whether the results of this study are broadly applicable (see Limitations), they can be used as the starting point for investigating classroom practice and pondering the stereotypes that students may carry with them into classrooms. For instance, undergraduates appear to have positive feelings about how professors organize and
understand the content they teach, and respect them overall, but they also appear to have more negative views of their abilities to relate to undergraduates and understand them. Conversely, undergraduates have some negative perceptions of GTAs' abilities to convey information instructionally, but they have very positive feelings about GTAs' abilities to interact with them. It is important to point out that these results do not seem to indicate that GTAs are less favored as instructors compared with professors; however, the stereotypes articulated by undergraduates almost surely impact the teaching and learning environment in undergraduate classrooms. Even if an individual GTA or professor does not adhere to the identified stereotypes, they are likely being compared with this typical perception, which may impact how students react to them in the classroom.

Our results are similar to those obtained by Park (2002) and Muzaka (2009) in which GTAs, staff, and undergraduates indicated that GTAs were understanding, approachable, laid-back, and nervous; have limited control and authority; and lack content knowledge and experience. Yet Park (2002) and Muzaka (2009) found that GTAs exhibit a youthful enthusiasm, while undergraduates from our study indicated that GTAs are less enthusiastic than professors. This may be a result of this study asking students to directly compare GTAs with professors, or because these data were collected solely from undergraduates. These differences may also be attributable to culture, since the work done by Park (2002) and Muzaka (2009) was completed in the United Kingdom, while our study took place in the United States. Our study also focused specifically on introductory courses in which laboratories function somewhat separately from lecture (taught by different instructors in different locations with different class sizes), whereas other institutions may have a different course structure. There are also different selection processes for GTAs at different universities, with some GTAs volunteering or being more enthusiastic about teaching, while GTAs at a research university may not be as enthusiastic or have the time or encouragement to embrace their teaching duties.

It is also now possible to compare the perceptions of undergraduates in this study with reflections of GTAs on their own teaching characteristics. Dudley (2009) documented instructor variables of GTAs from the perspective of GTAs. These GTAs expressed ideas such as: confusion with expectations, difficulty establishing boundaries, dealing with nervousness, expectation to be knowledgeable, and being uncertain where to start. The variables of nervousness and being uncertain about where to start were also identified as characteristics of GTAs by undergraduates in our study. The variables of confusion with expectations, difficulty establishing boundaries, and expectation to be knowledgeable identified by Dudley (2009) are similar to the lack of experience, control, and knowledge that undergraduates identified with GTAs in this study. These similarities in the perceptions of GTAs by GTAs themselves and undergraduates may be explained by the fact that GTAs see themselves as students who merely have teaching responsibilities (Park, 2002; Muzaka, 2009).

Proposed Explanations of Instructor Differences

The aspects that undergraduates perceive as different between GTAs and professors may have their origin in factors that are specific to the academic context, such as who has control over the curriculum and the status of the instructor (faculty or not faculty). When using these factors to explain the themes (delivery technique, classroom atmosphere, and relationship) it should be acknowledged that one factor may influence several aspects of instructor perception. Some of these factors will be highlighted below by providing literature that explains differences between student perceptions of GTAs and professors for each of the themes identified in this study.

One factor that may contribute to undergraduate perception of the instructor’s delivery technique is the curriculum. Professors typically have more control over the organization of the curriculum and classroom policies, while GTAs are typically given specific assignments within the curriculum to enact, with oftentimes little opportunity to alter or modify it. This was articulated in a study by Park and Ramos (2002), in which GTAs reported they had little autonomy or ownership over what they taught, but were merely “carrying out the job,” and a study by Muzaka (2009), in which GTA lack of control and authority over the curriculum led to a perception that students “see no point to us.” The perception of “control,” however, could also be due to professors having a greater student–teacher distance, due to status, age, and possibly greater confidence in the subject matter (Roach, 1997). For undergraduates in our study, these factors may have contributed to their feelings of their GTAs being hesitant and uncertain about what they were teaching and their professors being confident, knowledgeable, and organized.

The classroom atmosphere theme appears to be influenced by the perception of instructor behavior in the classroom. For instance, GTAs are perceived as being engaging and laid-back, while professors are more distant and formal. Even though the classroom itself was not perceived by students as a distinguishing variable, it could be that the size of the classroom affects classroom teaching behavior. GTAs often teach smaller sections (laboratories and discussion sections) in which students are able to interact with them on a one-on-one basis, while professors often have larger classes in which it is more difficult to interact with students individually (Dudley, 2009). Professors may also appear more distant due to knowledge level and age differences, which may limit the interactions they have with students (Anderson and Carta-Falsa, 2002). Undergraduates may also perceive that since GTAs are typically similar in age to them, they can better relate to their classroom experiences and explain things to them (Muzaka, 2009).

Certainly, this perception of engagement is a positive aspect of GTA instruction. Darby (2005) concluded that when GTAs are enthusiastic about the subject matter, students are more comfortable with the subject matter and their learning is better supported. Similarly, O’Neal et al. (2007) determined that GTA enthusiasm positively impacts student retention in the sciences. This may be attributed to engaged students being more likely to learn and retain knowledge (Umbach and Wawrzynski, 2005).

For the theme of relationship, instructor age may be a factor influencing student perception. GTAs are typically younger than faculty members, and thus could be perceived by the undergraduates as closer to their own age and therefore more approachable (Muzaka, 2009). GTAs and undergraduates also have similar experiences, because GTAs are often still taking graduate classes while they are teaching (Park and Ramos, 2002).
2002; Dudley, 2009). Thus, undergraduates may think that since GTAs are more familiar with academic demands, the pressures of deadlines, and workload in their own courses and research, they may be more understanding, approachable, and relatable when undergraduates express workload issues (Muzaka, 2009). Greater age, status, and confidence (Roach, 1997) may also be why undergraduates in this study afforded professors more respect than GTAs.

Limitations

One limitation of this study is the small sample size and the fact that it was conducted on a limited subset of volunteer participants in one discipline at one university. The sample was also greatly overrepresented by first-year, female, non-biology major undergraduates. This particular sample may have had limited experience with instructors of various titles (GTA vs. professor), and the respondents therefore may have been thinking of the one GTA/professor they had in college when responding. This study is also limited in that it was performed at a large southern research university with its own culture of instruction and curriculum that may not be present at the vast majority of other schools. Further, student demographics for each of the quantitative scenarios were not identical. As with many single-institution studies, the results cannot be generalized to all academic institutions; however, they can and should be used as the basis for additional investigations of GTA and professor instructional characteristics.

Recommendations for GTA Professional Development

The findings of this study suggest that undergraduates may have different perceptions of GTAs and professors with regard to several important instructional aspects. While additional investigations are certainly needed into why these factors are perceived differently by undergraduates, and whether these hold true at different institutions, these factors can be used in the meantime to help shape GTA professional development.

GTAs could be made aware of the results so they can better understand how undergraduates perceive them and come to know that most undergraduates do not understand the academic context of graduate students. GTA professional development should focus on keeping the positive aspects of GTA instructors (such as relatable, engaging, and approachable) while finding ways to decrease the perceived nervousness, uncertainty, and hesitancy of GTAs. For instance, universities can work to better prepare GTAs for the specific curriculum content they will be teaching. Marinovich et al. (1998) noted that GTA assignments are often made just prior to the semester, and suggested that if these assignments were earlier, GTAs might be able to better prepare for the courses they will be teaching. Faculty often know their course assignments months in advance and invest considerable amounts of time preparing course materials and syllabi; GTAs should be given the same advance preparation time, if possible.

Shannon et al. (1998) and Luft et al. (2004) stated that GTAs should know more about curriculum delivery in general, and that providing professional development sessions in which GTAs are given background to deliver specific curricula may make them feel more confident and knowledgeable in the classroom. In addition to training in delivering specific curricula, GTAs should be encouraged to take time to reflect on their teaching for the purposes of self-evaluation; this will allow them to more quickly identify their personal teaching style, which should better promote student learning (Schussler et al., 2008).

GTAs could also be coached on behaviors that help them strike a balance between informal and relaxed and being strict and having high standards, because the former may be beneficial in terms of engaging students, but it can also be problematic in terms of authority and respect (Muzaka, 2009). Roach (1997) suggested that attire impacts student perception of the instructor; instructors dressed in professional attire may be afforded more classroom control and authority. Thus, GTAs can be informed of the importance of proper attire when instructing students.

This study identified positive and negative aspects of both GTA and professor instruction from the perspective of undergraduates. Often, where one instructor was weaker in an aspect, the other was strong. This suggests that another mechanism of professional development would be giving more opportunities for GTAs and professors to teach collaboratively, which could help each gain in aspects that are perceived as weaknesses by undergraduates. For instance, a professor working with a GTA on a course could help the GTA develop a better understanding of how to be more confident and organized about the curriculum and how to set higher standards. To achieve this, lecture and laboratory courses could be co-taught by professors and GTAs, versus separating those duties, or in cases in which this is not possible, GTAs could help coordinate the curriculum for a course they have been assigned to teach.

These recommendations for GTA professional development focus on maintaining as many of the positive aspects of GTA teaching as possible, while simultaneously finding ways to decrease the negative perceptions expressed by undergraduates. The overall goal of these new programs would be to increase GTA confidence in their teaching assignments and undergraduate perception of their teaching abilities, which should result in a better learning environment for everyone.

CONCLUSIONS AND FUTURE DIRECTIONS

There were several factors that undergraduates in this study perceived as being different between professors and GTAs; this study documented these differences and made suggestions for how to use this information to potentially improve teaching and learning at universities. Additional research, however, could further refine these results. For instance, exploration into why undergraduates perceive these differences may clarify why first-year students in this study had similar viewpoints compared with upper-level students with more exposure to different instructor types. Other research could study actual GTAs and professors to see whether the differences identified in this study hold true in practice or at other institutions. It would also be interesting to study undergraduate student perception of instructors who deviate from stereotypical instructor variables and how this impacts student perception of the instructor. For example, a professor who is nervous or a GTA who is unapproachable may be perceived more negatively, because they do not adhere to student expectations for that type of instructor. Future studies should also clarify the terms undergraduate students used
in the study, for example, what exactly students mean when they say “strict” or “uncertain.” Careful studies such as these will shed additional light on the complicated instructional relationships among undergraduates, GTAs, and professors, which may help each group to better understand how to maximize teaching and learning in undergraduate courses in the future.

ACKNOWLEDGMENTS
We thank the Division of Biology and Department of Ecology and Evolutionary Biology at the University of Tennessee for funding this project. We also thank the instructors and students who participated, as well as M. L. Niemiller for reviewing previous drafts of the manuscript.

REFERENCES
Austin AE (2002). Preparing the next generation of faculty: graduate school as socialization to the academic career. J High Educ 73, 94–122.
Bolge RD (1995). Examination of Student Learning as a Function of Instructor Status (Full-Time vs. Part-Time) at Mercer County Community College. West Windsor, NJ: Mercer County Community College.
Students in an interdisciplinary undergraduate introductory course were required to complete a group video project focused on nutrition and healthy eating. A mixed-methods approach to data collection involved observing and rating video footage of group work sessions and individual and focus group interviews. These data were analyzed and used to evaluate the effectiveness of the assignment in light of two student learning outcomes and two student development outcomes at the University of Minnesota. Positive results support the continued inclusion of the project within the course, and recommend the assignment to other programs as a viable means of promoting both content learning and affective behavioral objectives.

INTRODUCTION

This article reports on the efforts of a biology professor at a large research university to address both science content learning objectives and behavioral and affective understandings for first-year students in an interdisciplinary course. The class assignment discussed here aims to provide experiences deep enough to transfer beyond a single class and assist students in developing skill sets useful to future endeavors both inside and outside academia. Helping students develop the capacity to know their own strengths, weaknesses, and motivations is a necessary part of their transformation into adults able to engage in “collaborative social relations with diverse others,” a skill employers now expect higher education to address (Baxter Magolda, 2008, pp. 269–270). Student assignments that promote multiple institutional goals are needed in light of new requirements that coursework address both specific content and developmental outcomes. By exploring the effectiveness of particular approaches to meet changing institutional needs, science instructors will be better prepared to draw upon a variety of strategies in a shifting academic climate.

Science educators have long valued a range of skills beyond a student’s content knowledge, and often require demonstration of laboratory procedures, proficiency in writing lab reports, and oral communication skills in presentations. Competencies associated with student development goals, however, represent a new domain for most instructors at the college level. Few science educators work in learning environments in which there has been overt direction to emphasize the need for interpersonal skills to better interact with diverse groups, yet we recognize that ability as valuable—if not essential—for success in professional science careers. Concerned instructors find ways to recognize the symbiotic relationship between these skills by promoting and combining learning and development outcomes in the same assignments. Providing opportunities for students to work in groups on assignments that prompt them to communicate ideas and promote both content and developmental goals is often recommended as a strategy to improve science, technology, engineering, and medical (STEM) education at the college level (Lord, 2001). Particularly in classes attended by nonscience majors, engaging students is important for success. Additionally, failing to create classroom environments in which students are encouraged to ask questions and feel comfortable expressing confusion contributes to many promising students’ deciding not to pursue a science major while in college (Seymour and Hewitt, 1997). In their discussion of a program designed to increase scientific literacy in a humanities-focused student body, McPhearson and colleagues (2008) emphasize the effectiveness of inquiry-based,
multidisciplinary approaches to teaching science content in a diverse class of incoming freshmen. Reforms designed to improve retention and address the problem of underrepresentation in science majors have also been presumed to benefit all students (Seymour, 2001).

In 1999, the National Research Council (NRC) called for science courses to become more authentic and to engage students in activities practiced by professional scientists. In the institutional context of this study, “authenticity” is interpreted as the need for students to develop skills to effectively collaborate with others in finding solutions to problems and creating a product in a team setting. The NRC specifically recognized the need to transform teaching and learning in institutions of higher education and recommended the exploration of STEM concepts “as practiced by scientists and engineers,” which includes the sharing and discussion of ideas (NRC, 1999). Additionally, the NRC report also called for introductory science courses to be more inclusive and to meet the needs of learners from diverse backgrounds, a goal shared by higher educational institutions concerned with retention and recruitment of diverse student bodies (and related to the establishment of student development and learning goals discussed in this article). Many other national policy documents (see A New Biology for the 21st Century [NRC, 2009]) recommend such behavioral goals, which cannot be readily achieved through the traditional lecture setting, and emphasize the need for alternative pedagogies to promote more collaborative science experiences. Designing a curriculum that optimizes student–student interaction to stimulate creative, critical thinking requires attention to the way in which group activities are structured, and also the type of questions that are explored collectively. Cooperative group learning has historically been recommended as a means to begin addressing developmental and academic outcomes that cannot be targeted through traditional lecture but, as Seymour (2005) notes, “…how to best infuse large university classes with more of the active and interactive learning methods remains a consistent challenge” (p. 1).

Cooperative group learning is an admittedly complex teaching and learning strategy with which many science instructors have limited experience. Grounded in the constructivist theory of learning (see Johnson and Johnson, 1989), this pedagogical approach posits that knowledge is fostered and organized through social interactions, such as those emphasized in cooperative group learning. Johnson and Johnson (1989) report that specific criteria must be met to achieve genuine cooperation among individuals within a group. For example, in order for groups to be truly cooperative, they must have some degree of positive interdependence among group members, and there must also be a strong individual accountability component in assignments. Group assignments that promote positive interdependence and individual accountability are part of creating more dynamic and interactive classrooms and thus initiate important steps toward the goals outlined by NRC and other invested groups. Research on the effectiveness of cooperative group learning is robust in terms of promoting both learning and development goals (Lord, 2001). The effectiveness of group learning has been shown in studies by Knight and Wood (2005), who found that increasing interactive and participatory activities significantly improved student learning; Eisen (1998), who found that group projects help engage students and improve content literacy; and DebBurman (2002), who showed that working in groups promotes developmental skills, such as critical thinking, communication, and social responsibility. In addition to being a heavily researched pedagogical strategy, cooperative learning has a “variety of positive and measurable outcomes on students at a variety of cognitive levels and in a variety of disciplines” (Tanner et al., 2003, p. 2).

In addition to changing pedagogical strategies, organizational documents promote modifying assessment practices. For example, the NRC (2003a) advocates the inclusion of both student learning and development outcomes in evaluation. More specifically, instructors are encouraged to move beyond a reliance on assessment of biological content knowledge through lecture exams, and even procedural skills shown in lab exams. Efforts to include skills and personal characteristics often documented in letters of recommendation, but not frequently assessed in traditional STEM courses (such as oral and written communication skills and the ability to work well with others) require broadening instructors’ understanding of assessment and class objectives. Instructor observation can be used to identify interpersonal communication process skills, such as the ability to interact with diverse groups of people. Dutson et al. (1997) also described how this observation simultaneously provides instructors the opportunity to inspect for more traditional learning outcomes, such as students’ application of content knowledge. Seymour (2001) includes the creation of assessment instruments to match reformed teaching strategies as a key to innovation in undergraduate science education.

Universities and colleges are now expected to do more than produce content specialists; we are to prepare students to function as productive and well-informed citizens (American College Personnel Association [ACPA] and National Association of Student Personnel Administrators [NASPA], 2004). To promote these goals, many universities now promote both learning outcomes and developmental outcomes. Learning outcomes have an extensive history in academics and typically state what knowledge should be known at the time of graduation (e.g., “ways of knowing” within a specific discipline). Comparatively, developmental outcomes are a new entity, and include such characteristics as the ability to work well in diverse environments and tolerance of ambiguity. NASPA and ACPA promote the “integrated use of all of higher education’s resources in the education and preparation of the whole student” and describe “learning” as a combined activity of academic and student development experiences (ACPA and NASPA, 2004). Baxter Magolda’s framework of constructive-developmentalism links learning conditions that embrace constructivist pedagogical approaches and understanding of learning with individual development processes (Baxter Magolda, 2000); such an understanding of the ways in which student growth occurs both intra- and interpersonally underlies the simultaneous pursuit of student development and learning outcomes. Furthermore, the interdisciplinary, cooperative approaches show great promise in improving student learning gains (as compared with traditional lecture style) and can serve as a more accessible entry point for nonscience majors to engage with and explore scientific investigation (McPhearson et al., 2008).
Many higher education institutions now apply constructivist theories of knowledge development to first-year student experiences; tracking student satisfaction and success is also an important application of efforts to evaluate courses for their incorporation of particular outcomes. Historically, freshman students at large universities have frequently attended large lecture courses that do not afford much individual attention. Recently, however, researchers such as Kuh (2008) and Tinto (2006–2007), have targeted the freshman experience as critical to establishing and maintaining student satisfaction and related to retention and graduation rates. Based on this research, many institutions have directed efforts toward improving the first-year experience by creating programs such as freshman seminars, extended freshman orientation, and common book experiences. Such ambitious goals of improving both instruction and student experience pose new challenges for instructors. In this paper, we offer our experiences at the University of Minnesota (UMN) and suggestions of how a similar approach could be applied at other institutions.

The following section of this article gives an overview of the history and current implementation of student learning and development outcomes at the heart of the study discussed here; this study aimed to explore the effectiveness of a first-year class project in supporting student progress toward selected objectives.

**Student Learning Outcomes and Student Development Outcomes at UMN and the First-Year Experience Initiative**

In the College of Education and Human Development (CEHD) at UMN, all incoming students participate in a First-Year Inquiry (FYI) course that promotes “multidisciplinary ways of knowing” (see Supplemental Material). Such courses are considered a “high-impact” activity in Kuh’s discussion of initiatives likely to promote undergraduate student success and engagement (Kuh, 2008). The FYI course was developed in 2007 by a team of administrators and instructors within the Department of Postsecondary Teaching and Learning, who incorporated two UMN student learning outcomes (SLOs; Can Communicate Effectively; Have Acquired Skills for Effective Citizenship and Lifelong Learning) and two student development outcomes (SDOs; Responsibility and Accountability; Appreciation of Differences) into the course proposal. These four outcomes address both the individual self-reflection necessary for critical thinking and the interpersonal skills required for effective social relations—a process referred to by Baxter Magolda (2008) and others as “self-authorship.” Additionally, these outcomes were aligned with the writing skills and multicultural focus of the college as a whole.

A set of SLOs and a separate set of SDOs were developed and published at UMN in an organized effort to stimulate the transformation of undergraduate education. Efforts began in May 2003, when the UMN’s Provost’s Council for Enhancing Student Learning (CESL) adopted a set of statements designed to guide the development of the outcomes. These statements called for the deliberate collection and analysis of evidence to determine student progress toward high standards, with the overall purpose of improving teaching and learning (personal communication, C. Murdoch, September 2010). Following these initial steps, seven SLOs were developed in the 2003–2004 school year by a curriculum-assessment working group of the Provost’s CESL. The seven SLOs were subsequently adopted by the university’s Faculty Senate as official policy. In 2007, the University Senate Education Policy Committee used a similar process to develop and adopt seven SDOs to complement the SLOs. Together, UMN views the SLOs and SDOs “entwined as critical elements of the student experience” (UMN, 2011a). Figure 1 contextualizes the role of SLOs and SDOs from an institutional, instructor, and student perspective. Table 1 displays a complete list of the SDOs, which “assist students to become lifelong learners and engage as effective citizens when they leave the University.”

![Figure 1. Contextual framework of student learning and development outcome implementation.](image)

<table>
<thead>
<tr>
<th>Table 1. UMN student development and learning outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>UMN SDOs</strong></td>
</tr>
<tr>
<td>Accountability and Responsibility&lt;sup&gt;+&lt;/sup&gt;</td>
</tr>
<tr>
<td>Independence and Interdependence</td>
</tr>
<tr>
<td>Goal Orientation</td>
</tr>
<tr>
<td>Self-Awareness</td>
</tr>
<tr>
<td>Resilience</td>
</tr>
<tr>
<td>Appreciation of Differences&lt;sup&gt;+&lt;/sup&gt;</td>
</tr>
<tr>
<td>Tolerance of Ambiguity</td>
</tr>
<tr>
<td>UMN SLOs</td>
</tr>
<tr>
<td>Can Identify, Define, and Solve Problems</td>
</tr>
<tr>
<td>Can Locate and Critically Evaluate Information</td>
</tr>
<tr>
<td>Have Mastered a Body of Knowledge and Mode of Inquiry</td>
</tr>
<tr>
<td>Understand Diverse Philosophies and Cultures within and across Societies</td>
</tr>
<tr>
<td>Can Communicate Effectively&lt;sup&gt;+&lt;/sup&gt;</td>
</tr>
<tr>
<td>Understand the Role of Creativity, Innovation, Discovery, and Expression across Disciplines</td>
</tr>
<tr>
<td>Have Acquired Skills for Effective Citizenship and Lifelong Learning&lt;sup&gt;+&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>+</sup>An outcome targeted in this study.
(UMN, 2011a), and the SLOs, which guide “what students know and can do” (UMN, 2011b), and indicates the four outcomes discussed in this article.

Currently, university policy directs instructors to incorporate the SLOs and SDOs into their course expectations but allows flexibility in the determination of how to do so. Curriculum committees within UMN use SLOs and SDOs in the process of developing and approving new courses, and also for periodic evaluation of existing courses. Instructors are not currently required to include all of the SLOs and SDOs in their class designs; ongoing evaluative projects, such as this study, will help inform the use of these outcomes in future course syllabi and assessments.

Six different sections of the FYI course have been offered each Fall semester since the course was initially proposed and approved by the college’s curriculum committee. Each Fall semester, ~450 students are divided into six different FYI classes, each of which is team-taught by three instructors (~75 students per class section). Groups of 75 students meet with all three instructors for 2 h/wk during the semester, and also meet in groups of 25 students with one instructor for 2 h/wk. The six different sections of the FYI course have a wide variety of topics and titles (e.g., Ecological Hotspots, Energy...Illuminated), but all must use a common set of assignments. For example, all sections must meet the requirements for a “writing-intensive” class, and thus must require at least 12 pages of revised text. A significant percentage of the course grade is to be based on students’ writing abilities. All FYI students must participate in an end-of-semester group project in which students build on the course concepts to collectively answer the question “How can one person make a difference?” The student projects are publicly presented at the end of Fall semester at a public capstone showcase that features all 450 students showing and viewing projects.

For 3 yr, the Food for Thought and Action FYI course has used a group video assignment to fulfill the requirements of the group capstone project. In 2010, a small grant was secured from UMN’s Academy of Distinguished Teachers to conduct an evaluation of the group video project in light of UMN’s SLOs and SDOs. Funds were used to hire a graduate student to coordinate the study independent of the day-to-day operation of the course, with results of the evaluation intended to be used in instructor and department decisions to drop or modify the capstone video assignment, or to retain and publicize it as a viable mechanism to promote a practical cooperative learning assignment promoting student behaviors that meet UMN’s SLOs and SDOs. In keeping with institutional and departmental goals of promoting SDOs, as well as requirements to assess student progress toward learning outcomes (see Figure 1), FYI instructors endeavor to create class assignments that provide an opportunity for both authentic science experiences and collaboration with others.

The purpose of the research project discussed here was to determine how a cooperative class assignment could be used to ascertain student progress toward both SLOs and SDOs. We conducted a qualitative evaluation in which student behaviors were observed and rated to identify specific instances consistent with the objectives of the selected SLOs and SDOs. The data from this evaluation were used to construct an understanding of how successful the cooperative group assignment was in meeting the instructional goals—both content learning and affective personal development—of the video activity.

**METHODS**

**Course Design: Food for Thought and Action**

The three instructors of the Food for Thought and Action class included a biologist, a writing instructor/lawyer, and a social scientist. The curriculum for the course was based on three different texts: *Food Inc.* (Weber, 2009) and two books from Michael Pollan—*In Defense of Food* (2008) and *Food Rules* (2009). Daily lessons in the course were developed by each of the three instructors and ranged from understanding the requirements for the U.S. Department of Agriculture’s “Organic” label (led by the writing instructor/lawyer), exploring the sociological dynamics involved in local farmers’ markets (led by the social scientist), and examining the biology of atherosclerosis, diabetes, and other diseases related to obesity (led by the biologist). The instructional strategies used in the large class meetings (75 students with all three instructors) were largely influenced by the classroom design—a 90-student active-learning room that featured 10 round tables. The room’s structure intentionally makes long lectures impractical but facilitates group interaction. In the Food for Thought and Action class, a typical lesson involves a short presentation by the instructor followed by a student group activity, and then a follow-up large-group discussion and summary. Assessment of student performance in the course involved several writing assignments (including several short papers and a traditional term paper), midterm and final exams, and the final group video project.

**Capstone Assignment Design**

The capstone assignment was introduced during the 10th week of the semester, when instructors assigned students to groups. The three professors used different selection methods, but each used an intentional strategy to create groups of four students. One professor focused on how well the students had done on the midterm exam, matching students who performed at different levels of academic proficiency; another created groups with a mixture of previously demonstrated social participation and personalities; and the third professor aimed to balance gender and students from different racial backgrounds.

1. The $70 million Science Teaching and Student Services (STSS) Building at UMN was completed in the Summer of 2010 and features “active-learning classrooms” that were designed based on Robert Beichner’s “SCALE-UP” project (http://scaleup.ncsu.edu). Nine round tables that promote student–student interaction and greatly enhance all cooperative group endeavors are the central features of each room. Large flat-screen computer monitors and access to ample whiteboard space accompany the tables. A lecture station is located at the middle of each room and includes a number of audiovisual options (computer, DVD, etc.), but the room is not designed for extended lectures. Rather, the setup specifically accommodates and encourages group work. (Pictures of the University of Minnesota’s active-learning classroom can be viewed at: www.classroom.umn.edu/projects/alc.html.)

2. Within the FYI course, a final project was required of all students. This final course project was called the capstone project. The term “capstone” has also been used to define a large-scale project at the end of an academic program. Here, however, it is used to describe the final project in a course.
and ethnic backgrounds in each group. The assignment and description of the video project was then explained to the students. (See Supplemental Material for project description and class syllabus.) The stated goal of the assignment was to “create a 30 to 60 second public service announcement (PSA) video focusing on food and targeting a specific audience.” More specifically, videos were to focus on one or two specific “food rules,” as outlined in Michael Pollan’s book Food Rules (2009).

The project description provided recommendations for the production process, such as guidelines for managing the initial brainstorming sessions. The students were encouraged to draft a storyboard to assist in developing and connecting ideas when initiating the project. The final product was to be filmed using a digital camera and edited using computer applications, such as iMovie or Movie Maker. Cameras, computers, and video-editing software were made available to students through the university library system, though many groups elected to use their own equipment. Students completed about one-fourth of the project during regular class hours, with the remaining hours scheduled according to group members’ availability (and without instructor oversight). Students had 5 wk to complete the project, and all videos were presented to the class at the culmination of the semester.

**Participants in the Research Study**

All students in the course were invited to voluntarily participate in the research study prior to the beginning of the capstone project. Students indicated their willingness to participate by signing a consent form that had been approved by the university’s institutional review board office. After groups were initially formed by instructors, a sample was selected by identifying those groups in which all four members had consented to participate in the research study. A total of six groups were selected: two from each instructor’s list. The groups were not notified that they would be recorded on video until the first day of the project, during their initial group meeting. Demographic data for the Food for Thought class showed it to be a relatively representative sample of students enrolled in CEHD, a college that is generally more diverse than the university as a whole (Table 2). Additionally, students in the course represented a cross-section of programs, such as university honors, TRIO (federally funded programs that support college opportunities for students from disadvantaged backgrounds; Council for Opportunity in Education), Access to Success (ATS), and Commanding English (CE; Table 3).

**Table 2.** Comparison of race/ethnicity of students enrolled in the Food for Thought and Action FYI course, freshman students in the CEHD, and freshman students at UMN

<table>
<thead>
<tr>
<th>Race/Ethnicity</th>
<th>FYI class (n = 84)</th>
<th>CEHD freshmen (n = 447)</th>
<th>UMN freshmen (n = 4876)</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>61%</td>
<td>61%</td>
<td>78%</td>
</tr>
<tr>
<td>Asian</td>
<td>19%</td>
<td>16%</td>
<td>9%</td>
</tr>
<tr>
<td>Black</td>
<td>14%</td>
<td>15%</td>
<td>2%</td>
</tr>
<tr>
<td>American Indian</td>
<td>4%</td>
<td>2%</td>
<td>1%</td>
</tr>
<tr>
<td>Hispanic</td>
<td>1%</td>
<td>2%</td>
<td>3%</td>
</tr>
<tr>
<td>Hawaiian/Pacific Islander</td>
<td>0%</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>Total</td>
<td>99%</td>
<td>97%</td>
<td>94%</td>
</tr>
</tbody>
</table>

*Not all students chose to identify ethnicity.

**Evaluation Design and Data Collection**

The research design utilized in this study includes both open-ended and close-ended approaches to data collection, in keeping with Cresswell’s description of concurrent mixed-method strategies (Cresswell, 2009). Using a structured rubric to track frequency data, videographers collected measures of student behaviors, which were analyzed quantitatively, while observation and emergent interview design allowed for qualitative investigation. The data-collection techniques used in this study provide a degree of triangulation aimed at establishing validity of the conclusions drawn from the evaluation. In particular, data were gathered through observing and recording video of group meetings, conducting individual interviews with representative students, and interviewing a focus group composed of students from each of the groups. Qualitative research methods allow for the investigation of a particular phenomenon or people’s experience of it within a defined context (Patton, 2001); educational settings and group dynamics are well suited to investigation using this approach. Following Tashakkori and Teddlie’s framework for mixed methods (Tashakkori and Teddlie, 2002), this study is primarily qualitative in nature, with research objectives aimed at building understanding of a particular experience supported by the collection of quantitative data (i.e., QUAL-quant). A combination of statistical and text analyses were used to investigate student perceptions and experiences, in keeping with the pragmatic methods advocated by many mixed-methods proponents (Johnson and Onwuegbuzie, 2004).

Evaluation as it is understood in this study follows Patton’s (2001) definition of this mode of investigation as “the systematic collection of information about the activities,
characteristics, and results of programs to (1) make judgments about the program, (2) improve or further develop program effectiveness, (3) inform decisions about future programming, and (4) increase understanding” (p. 39). Following Merriam (1998), interviews serve an important interpretive function, because they allow individuals to explain internal thought processes and explain personal experiences. Focus groups can allow for more open-ended questioning and for participants to interact with one another and build upon others’ answers, and are a way to encourage participants to express their opinions and perceptions of a shared experience (Krueger and Casey, 2009). The evaluation of the group capstone video project was conducted using the four SLOs and SDOs incorporated in the design of the FYI course and used as a guide to measure student progress in keeping with the university’s goals.

As described earlier in this paper, the four SLOs and SDOs were identified by a departmental curriculum committee and associated objectives were specified for each outcome. This list of objectives was intended to provide a mechanism through which data generated from observing student behaviors could be analyzed. A program evaluation specialist developed the particular wording of the objectives, and they were previously used on other projects. The four SDOs, as well as the objectives for each outcome are listed in Table 4.

Videographers for the project were selected from a group of graduate students from the Counseling and Student Personnel Psychology program within CEHD, who were trained in recording techniques by a university media specialist and instructed in how to record, secure, and share their video data.

**Table 4.** Student development and learning outcomes and associated objectives

<table>
<thead>
<tr>
<th>SDO 1: Accountability and Responsibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group members decide and carry out specified roles.</td>
</tr>
<tr>
<td>Group meetings are set for times that meet all group members’ schedules.</td>
</tr>
<tr>
<td>Group members show up for meetings.</td>
</tr>
<tr>
<td>Group members create a feasible timeline that includes who will complete specific tasks.</td>
</tr>
<tr>
<td>SDO 2: Appreciation of Differences</td>
</tr>
<tr>
<td>All group members invite and listen to one another’s ideas and perspectives.</td>
</tr>
<tr>
<td>All group members are given time to offer support and explanation for their ideas and/or perspectives.</td>
</tr>
<tr>
<td>Final decisions relative to video content are negotiated and agreed upon by all group members.</td>
</tr>
<tr>
<td>SLO 1: Can Communicate Effectively</td>
</tr>
<tr>
<td>Group members discuss and decide upon the audience for their video.</td>
</tr>
<tr>
<td>Group members communicate effectively with one another.</td>
</tr>
<tr>
<td>Group members develop together a storyboard that includes how the video message will be conveyed to a selected audience.</td>
</tr>
<tr>
<td>Actors use eye contact, confident and audible voice tone, and effective language throughout video presentation.</td>
</tr>
<tr>
<td>SLO 2: Have Acquired Skills for Effective Citizenship and Lifelong Learning</td>
</tr>
<tr>
<td>Group members bring at least two different disciplinary perspectives to their chosen topic.</td>
</tr>
<tr>
<td>Group members articulate how they envision their video message impacting the intended audience.</td>
</tr>
<tr>
<td>Group members demonstrate awareness of which tasks will benefit from several people working together.</td>
</tr>
</tbody>
</table>

Each group was recorded by the same individual throughout the project. Recordings of 25–30 min were made of each of the group’s meetings. The first recording took place during a regular classroom session, immediately after the students were assigned to groups. The second recording also took place during a regular classroom session, and occurred as groups set up storyboards and developed scripts for their final videos. The third took place outside regular classroom time and documented groups filming and/or editing their final products. All six groups were taped for the first and second sessions, but only four of the six groups were taped during the third session, due to scheduling conflicts.

**Scoring Rubric and Assessment**

Scoring rubrics were developed for each of the three recording sessions (see Table 5). Each rubric contained the four SLOs and SDOs, with two or three identified objectives for each outcome. Specific objectives used to evaluate each session were defined by the nature of the student tasks involved during each meeting. For example, within the SLO Accountability and Responsibility, one objective was “Group members carry out decided-upon roles and continue to decide what each member is responsible for throughout the remainder of the process.” This objective was germane to the beginning stages of the project and was used to rate sessions 1 and 2, but was not included in the rubric for session 3.

Two individuals rated each video session. “Rater 1” was the coordinating graduate student, who rated all three sessions for all six groups, and “Rater 2” was the videographer for each specific group. The scoring rubric used a Yes/No (i.e., forced-choice) system to identify whether or not specific objectives were met by the behaviors of the group during the recorded session. For example, regarding the objective “Group meetings are set for times that meet all group members’ schedules,” a Yes was marked if the group was observed verbally committing to future meeting times during the video session, and a No was marked if this did not occur. No timestamp comparisons were used while rating video sessions, and thus it was possible for the two raters to agree that a group did engage in a specific behavior related to an outcome, but identify different times within the session when they considered the criterion to have been met.

Student behaviors from the first group meeting were rated based on nine objectives (within the four SLOs/SDOs) and six groups, and thus a maximum of 54 Yes markings was possible per rater. Rater 1 identified 30 instances related to the objectives, whereas the six videographers identified 34 instances. Rating student behaviors for the six groups during the second group meeting (in which students made storyboards and specific plans for their videos) involved 10 objectives and thus a maximum of 60 possible Yes markings per rater. Rater 1
identified 40 instances related to objectives during this session, whereas the six videographers identified 39 instances. Eight objectives were used in rating the third student session, when students were filming and/or editing their projects. Only four of six groups were taped during this session, and thus the maximum number of possible Yes markings per rater was 32. Both Rater 1 and the four videographers identified 29 such instances while rating the third session. Out of a maximum 146 possibilities, Rater 1 identified 99 instances of student behaviors that were identified as meeting specific objectives within the scoring rubric, and the six videographers identified 102 instances. Based on the differences in rating, interrater reliability between the lead graduate student and the six videographers was determined to be ~98%.

**Student Interviews**

Individual and group interviews were conducted after completion of the class video project. Six students were interviewed individually, and six others met collectively for a focus group interview. Each of the six study groups had one participant in the individual interviews and one member in the focus group. Selection of students was based on schedule availability, (i.e., convenience sample). Both individual and group interviews were conducted using an open-ended, six-question protocol; two questions were the same for each group and four were intended to query students about events and opinions particular to their groups’ experiences. Individual interviews ranged in length from 10 to 15 min, and the focus group interview lasted 21 min. Both the individual and focus group interviews were recorded and transcribed. Specific student quotes that were germane to the analysis of the four SLOs/SDOs were extracted from the transcripts.

The groups included in this study were representative of students enrolled in the larger class in terms of gender, race, ethnicity, and English language proficiency. As this was an introductory course during the first semester of the students’ first year of college, students had similar levels of (un)familiarity with each other and with college-level course expectations. The data show that student groups generally made similar progress toward attainment of desired development and learning outcomes. The following section looks further at individual student experiences and examines nuances between group dynamics and individual reactions.

**RESULTS AND DISCUSSION**

The aforementioned study design was used to determine whether observed student behaviors indicated that targeted student development and learning outcomes were addressed through participation in this project. Following the evaluative approach described above, six student groups were recorded three times each throughout the course of the video project. After the project was completed, six individual students were interviewed about their experiences and a different set of six students participated in a focus group interview. Using the established rubric to assess recordings of group interactions, frequency measures were made to ascertain to what extent groups met the targeted development and learning outcomes.

Transcript data from interviews show students developing the ability to work through problems with classmates and discussing the course’s impact on how they may participate in cooperative group activities in the future. Students demonstrated the ability to assign individual roles within groups, while also entertaining multiple perspectives within discussions to meet the objectives of the video assignment.

Evidence that student behaviors were consistent with the goals of the selected learning and development outcomes was found in both the quantitative data (the frequency with which group members were observed to have met the objectives for each outcome) and the qualitative data (student responses in interviews and focus groups). This section summarizes analysis of video recordings and of narrative data gathered about each of the four learning and development outcomes, relying heavily on students’ own words to express what was learned.

**Data Related to SDOs**

**SDO 1: Accountability and Responsibility.** Coding for SDO 1 (Accountability and Responsibility) was based on objectives that identified behaviors related to assigning and carrying out specific roles, scheduling, developing a timeline, and showing up for meetings. Data for this outcome (Table 6) show 50% (for both Rater 1 and Rater 2) of the objectives met during the first recording, 72% (for both raters) met during the second recording, and 88% (for both raters) met during the third recording.

The data suggest that students in some groups initially struggled with scheduling demands and task organization and distribution. Group members found it easier to build consensus, however, as they became more comfortable with one another and more engaged in the project. For example, one student recalled:

> It was easier to meet at the end because I think everyone really wanted to get this done, and really wanted to do a good job, but I think it was harder at the beginning because we didn’t have any ideas and we did not know what to expect.

The above response also indicates that students were indeed interested in producing a quality final video, not in merely completing the assignment. Some groups immediately prioritized meetings and felt encouraged by one another’s commitment. A member of one such group reported that:

> It was a very great experience. Everybody was on time and everybody was willing to do something. Everybody did their part on time without missing anything.

In other groups, however, peer accountability took longer to develop. For example:

> The hardest part with everything was with our schedules. We had a lot of conflicts, and finally we said, “This is what we are doing and everything else is put aside.”

If we wouldn’t have done that, I don’t think we would have gotten as much done as fast.

Results here are also consistent with Johnson *et al.* (1991) and others who advocate groups staying together for more than one or two class sessions so that interpersonal dynamics can develop. This development outcome also highlights

---

3As mentioned previously, only four of the six groups were available for the final recording session.
Table 6. Scoring data for SDO 1: Accountability and Responsibility

<table>
<thead>
<tr>
<th>Taping 1: first group meeting (n = 6 groups)</th>
<th>Taping 2: storyboard/script development (n = 6 groups)</th>
<th>Taping 3: filming/editing (n = 4 groups)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student objectives</td>
<td>Rater 1</td>
<td>Rater 2</td>
</tr>
<tr>
<td>1. Group members decide and carry out specified roles.</td>
<td>33% (2 out of 6)</td>
<td>33% (2 out of 6)</td>
</tr>
<tr>
<td></td>
<td>67% (4 out of 6)</td>
<td>67% (4 out of 6)</td>
</tr>
<tr>
<td></td>
<td>100% (4 out of 4)</td>
<td>100% (4 out of 4)</td>
</tr>
<tr>
<td>2. Group meetings are set for times that meet all group member’s schedules.</td>
<td>67% (4 out of 6)</td>
<td>67% (4 out of 6)</td>
</tr>
<tr>
<td></td>
<td>83% (5 out of 6)</td>
<td>83% (5 out of 6)</td>
</tr>
<tr>
<td></td>
<td>——</td>
<td>——</td>
</tr>
<tr>
<td>3. Group members create a feasible timeline that includes who does what, as well as when parts of the video are due.</td>
<td>50% (3 out of 6)</td>
<td>50% (3 out of 6)</td>
</tr>
<tr>
<td></td>
<td>67% (4 out of 6)</td>
<td>67% (4 out of 6)</td>
</tr>
<tr>
<td></td>
<td>——</td>
<td>——</td>
</tr>
<tr>
<td>4. Group members show up for meetings</td>
<td>——</td>
<td>——</td>
</tr>
<tr>
<td>Percentage of Yes codes</td>
<td>50% (9 out of 18)</td>
<td>50% (9 out of 18)</td>
</tr>
<tr>
<td></td>
<td>72% (13 out of 18)</td>
<td>72% (13 out of 18)</td>
</tr>
<tr>
<td></td>
<td>88% (7 out of 8)</td>
<td>88% (7 out of 8)</td>
</tr>
</tbody>
</table>

The way in which group accountability and responsibility was reliant on individual accountability and responsibility, which is a necessary criterion for cooperative group learning (Johnson et al. 1991). Overall, most of the groups evaluated were observed to demonstrate a high degree of interdependence and to coordinate their efforts in a manner in which members held each other accountable for the final project.

**SDO 2: Appreciation of Differences**

Coding for the SDO 2 (Appreciation of Differences) was based on objectives that identified behaviors related to listening to the ideas of others, supporting and encouraging fellow group members, and coming to an agreement about the quality of the final project. Data for this outcome (Table 7) show 75% and 67% (Rater 1 and Rater 2, respectively) of the objectives met during the first recording, 92% and 83% (Rater 1 and Rater 2, respectively) during the second recording, and 100% for both raters during the third recording.

The researchers involved in this project concluded that many of the most powerful student reactions to the video project were related to this development outcome. Students reflected upon the experience as one that introduced them to new ideas, new people, and new ways of behaving in a classroom setting. Although cross-cultural sensitivity and open-mindedness can take more than a semester to develop, such experiences are valuable in helping students confront internalized stereotypes and preconceptions (Hlyva and Schuh, 2004).

Similar to the Accountability and Responsibility SDO, students reported understanding how their individual preferences and experiences related to the overall group interaction:

I like working alone, but this was my first group project in college. I felt like this introduced me to the college world working with three other people who you do not know very well. I was really quiet at first. As we worked through, I was opening up and going outside of my box.

Table 7. Scoring data for SDO 2: Appreciation of Differences

<table>
<thead>
<tr>
<th>Taping 1: first group meeting (n = 6 groups)</th>
<th>Taping 2: storyboard/script development (n = 6 groups)</th>
<th>Taping 3: filming/editing (n = 4 groups)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student objectives</td>
<td>Rater 1</td>
<td>Rater 2</td>
</tr>
<tr>
<td>1. All group members invite and listen to each other’s ideas and perspectives.</td>
<td>67% (4 out of 6)</td>
<td>67% (4 out of 6)</td>
</tr>
<tr>
<td></td>
<td>83% (5 out of 6)</td>
<td>83% (5 out of 6)</td>
</tr>
<tr>
<td></td>
<td>——</td>
<td>——</td>
</tr>
<tr>
<td>2. All group members are given time to offer support and explanation for their ideas and/or perspectives.</td>
<td>83% (5 out of 6)</td>
<td>67% (4 out of 6)</td>
</tr>
<tr>
<td></td>
<td>100% (6 out of 6)</td>
<td>83% (5 out of 6)</td>
</tr>
<tr>
<td></td>
<td>100% (4 out of 4)</td>
<td>100% (4 out of 4)</td>
</tr>
<tr>
<td>3. Final decisions relative to video content are negotiated and agreed upon by all group members.</td>
<td>——</td>
<td>——</td>
</tr>
<tr>
<td>Percentage of Yes codes</td>
<td>75% (9 out of 12)</td>
<td>67% (8 out of 12)</td>
</tr>
<tr>
<td></td>
<td>92% (11 out of 12)</td>
<td>83% (10 out of 12)</td>
</tr>
<tr>
<td></td>
<td>100% (8 out of 8)</td>
<td>100% (8 out of 8)</td>
</tr>
</tbody>
</table>
Table 8. Scoring data for SLO 1: Can Communicate Effectively

<table>
<thead>
<tr>
<th>Student objectives</th>
<th>Taping 1: first group meeting (n = 6 groups)</th>
<th>Taping 2: storyboard/script development (n = 6 groups)</th>
<th>Taping 3: filming/editing (n = 4 groups)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rater 1</td>
<td>Rater 2</td>
<td>Rater 1</td>
</tr>
<tr>
<td>1. Group members communicate effectively with each other.</td>
<td>17% (1 out of 6)</td>
<td>33% (2 out of 6)</td>
<td>83% (5 out of 6)</td>
</tr>
<tr>
<td>2. Group members discuss and decide the audience for video.</td>
<td>83% (5 out of 6)</td>
<td>100% (6 out of 6)</td>
<td>—–</td>
</tr>
<tr>
<td>3. Group members develop a storyboard that includes how their message will be conveyed.</td>
<td>—–</td>
<td>—–</td>
<td>67% (4 out of 6)</td>
</tr>
<tr>
<td>4. Actors use eye contact, confident and audible voice tone, and effective language throughout video presentation.</td>
<td>—–</td>
<td>—–</td>
<td>—–</td>
</tr>
<tr>
<td>Percentage of Yes codes</td>
<td>50% (6 out of 12)</td>
<td>67% (8 out of 12)</td>
<td>75% (9 out of 12)</td>
</tr>
</tbody>
</table>

As discussed earlier, a wide range of students from diverse backgrounds were enrolled in the FYI course, and many described the experience as one of learning how to interact with others in a new way. The need to be able to adapt one’s expectations to accommodate the ideas of others was frequently expressed; one student stated, for instance:

One thing that I would take away from this project was to keep an open mind when going into a project like this. An open mind with your group members, an open mind with the ideas people have, an open mind with how the filming process is going to go and how the editing is going to go. I guess I really learned what it means to be part of a group and to work cohesively together to create this great big awesome final product that we can be proud of.

Awareness of metacognitive processes was expressed most overtly in student responses to interview questions related to this outcome. As Baxter Magolda (2000) has noted, “the cognitive dimensions of self-authorship are intertwined with the interpersonal dimension” (p. 11), showing this link between this outcome and others investigated in this study. When asked to explain one of the most significant things he learned from the experience of working on the group project, one student described:

Working with somebody and having different opinions. They told you their side of the story and why they think that. [And I thought:] “OK, I did not think about that.”

Student responses also indicated that having a structured project guiding their group interactions made it easier to begin to get to know one another, but that once they felt more comfortable, the project itself became more enjoyable. In interviews with members of at least two groups, the dynamic appeared to switch from a task-oriented focus to one of member interaction over the course of the project.

Data Related to SLOs

SLO 1: Can Communicate Effectively. Coding for SLO 1 (Can Communicate Effectively) was based on objectives that identified student behaviors related to identification of a target audience for the video, development of a functional storyboard, and actors’ abilities to communicate within the final video. Data for this outcome (Table 8) show 50% and 67% (Rater 1 and Rater 2, respectively) of objectives met during the first recording, 75% (for both raters) during the second recording, and 100% (for both raters) for the third recording.

To investigate how well students were able to use strategies to encourage effective communication within their groups, representatives were asked to explain how decisions were made and how different ideas were considered. The selection of an audience for the public service component of the final video was an issue of considerable discussion in groups. When asked to explain why her group chose to create a video highlighting dining hall food options, one student said:

We thought that because of the problems with obesity now, it would be good to get our point to college students who can be the ones who change the obesity problem of the future.

The way in which groups decided to present their message also demonstrated an awareness of different communication styles and effectiveness. One group appealed to logic, for example:

We thought that we can’t change the older people, so we thought we would target the younger kids. We showed the younger kid going into the store, and then choosing which cereal is right for them, and going back and thinking, “OK, this is what I have.” We gave them reason[s] why this one is good and why this one is bad.

In contrast, another group relied on emotion:

If parents were to watch our video, they would definitely see our message from our skit. When you watch it, it shows “Why would you ever put soda into a sippy cup?” That goes right through their teeth. Their teeth
are going to rot. The kid is going to be malnourished. It is going to be terrible. So when you watch it, you think, “Why would a parent ever do that?” I definitely think that we got our point across well.

Communication skills were demonstrated by group members’ interactions with one another during classes and meetings, as well as in the way in which the final video transmitted each group’s desired message. Despite the short length (30–60 s), the videos were informative and entertaining and designed to reach a variety of audiences. (Example student video projects can be viewed at: http://msjensen.cehd.umn.edu/student-videos/food.asp and in this article’s Supplemental Material.)

SLO 2: Have Acquired Skills for Effective Citizenship and Lifelong Learning. Coding for SLO 2 (Have Acquired Skills for Effective Citizenship and Lifelong Learning) focused on the identification of skills that went beyond the requirements of the course, behaviors related to the potential impact of the video on the target audience, using multiple perspectives to convey a message, and awareness of group cooperation. Data for this outcome (Table 9) show 50% and 75% (Rater 1 and Rater 2, respectively) for the objectives met during the first recording, 39% (for both raters) during the second recording, and 75% (for both raters) for the third recording.

Although broad in possible meaning, demonstration of effective citizenship and lifelong learning skills was defined here as combining information from multiple sources, collaborating on a variety of tasks, and applying knowledge to new contexts. Some students were apprehensive at the beginning of the project, due to the required cooperative nature of the assignment and unfamiliarity with other group members, but gained new appreciation for the learning process throughout the semester. One student described her experience as:

To be honest, at the beginning of the project when we found out who our group members were, I did not really know any of them. I was kind of [wary] as to how it would all pan out. To be honest, I was pretty negative towards who my group members were because I did not know them. For me, personally, I have really high expectations of myself. And when I do projects like this, I really want them to turn out really well because that is who I am. But after talking to my group members, and brainstorming for ideas, and being on the same page, it worked really well. I take back everything that I thought at the beginning of the project, because I think that is what I learned out of working with people that you don’t necessarily feel comfortable with, finding a way to make it happen.

The necessity of building interpersonal trust and understanding in order to work together was acknowledged by students as a key part of being able to complete a group task that required combining multiple perspectives and ideas. As recounted by one student:

First of all, I did not know anybody, though I had class with them. To do this project, we had to know each other very well, and do work together. To start out with not knowing your group, and to come out with one final idea—that was a good experience—it was fun.

Students also expressed being intrinsically motivated by the process of working with other students rather than focused on the grade they would receive for the assignment:

The good thing was, when we met up for the video, we tried to have fun. We didn’t say, “Oh this is a project, we need to be all serious.” We had fun. We were going to go out there and try our best at it and not think about the grading process. I met a lot of new people. This was one of the first classes where I felt welcomed to the university. I am going to miss it.

The ways in which students were able to create a group space in which multiple perspectives were discussed and considered showed development of collaborative process skills and respect for others. Connections between citizenship and academic achievement were clearly identified by students upon reflection of their experiences.

Study Limitations

This study documented the ways in which students negotiated the inclusion of two SLOs and two SDOs in a single course project. As these outcomes are not

Table 9. Scoring data for SLO 2: Have Acquired Skills for Effective Citizenship and Lifelong Learning

<table>
<thead>
<tr>
<th>Student objectives</th>
<th>Taping 1: first group meeting (n = 6 groups)</th>
<th>Taping 2: storyboard/script development (n = 6 groups)</th>
<th>Taping 3: filming/editing (n = 4 groups)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rater 1</td>
<td>Rater 2</td>
<td>Rater 1</td>
</tr>
<tr>
<td>1. Group members articulate how they envision their video message impacting their intended audience.</td>
<td>83% (5 out of 6)</td>
<td>83% (5 out of 6)</td>
<td>50% (3 out of 6)</td>
</tr>
<tr>
<td>2. Group members demonstrate awareness of which tasks will benefit from several people working together.</td>
<td>17% (1 out of 6)</td>
<td>67% (4 out of 6)</td>
<td>33% (2 out of 6)</td>
</tr>
<tr>
<td>3. Group members bring in at least two different disciplinary perspectives on their chosen topic.</td>
<td>——</td>
<td>——</td>
<td>33% (2 out of 6)</td>
</tr>
<tr>
<td>Percentage of Yes codes</td>
<td>50% (6 out of 12)</td>
<td>75% (9 out of 12)</td>
<td>39% (7 out of 18)</td>
</tr>
</tbody>
</table>
department-specific, the rating rubrics used in this evaluation address somewhat broad categories of student behaviors and responses. Despite the difficulty in measuring abstract constructs such as Appreciation of Differences, the high degree of interrater reliability among researcher observations shows attention to consistency in the rating process. It is important to note, however, that the observed behaviors in this study all occurred within intentionally organized small groups of students that were maintained throughout the project. The results are therefore limited to instances that took place in the context of the video project and that were observed to relate to the four outcomes. Although student reflections suggest we can be cautiously optimistic that participants may transfer their learning beyond this experience, no specific claims can be made that these behaviors are transferable to new groups or different courses. Our hope, of course, is that individual gains have been made that can be applied to new group endeavors. This investigation therefore uses qualitative approaches to explore a particular phenomenon in depth, while acknowledging limitations of generalizability to a larger context. All conclusions reached from this evaluation are specific to the goals chosen as indicators of student progress toward the learning and development outcomes included in this particular course capstone project.

Furthermore, the implementation of such assignments requires previous instructor effort to prepare students to complete a final group project. Results of this evaluation are specific to the unique interdisciplinary class design and physical space (active-learning classroom); applying this approach to dissimilar settings likely requires some adaptation. Despite these limitations, the data collected show promising results that are encouraging for instructors and university administrators interested in incorporating multiple objectives into first-year undergraduate courses.

CONCLUSION

Kuh (2008) describes two key goals of collaborative learning as “learning to work and solve problems in the company of others, and sharpening one’s own understanding by listening seriously to the insights of others, especially those with different backgrounds and life experiences” (p. 20). Toward the goal of creating a campus culture that promotes student success and engagement, faculty can take advantage of the importance of peer influence by promoting student interaction through collaborative activities (Kuh et al., 2008). Additionally, many national organizations focused on improving higher education highlight the importance of adopting student-centered approaches for many reasons, including the need for effort to adapt to the increasing diversity of student populations (ACPA and NASPA, 2004). Understanding that classroom instruction is linked to student engagement and retention places additional importance on transforming pedagogical practice (Tinto, 2006–2007). Historically, science professors have focused on knowledge gains (learning outcomes) when assessing and evaluating students, assignments, courses, and even programs. Appreciation of differences, accountability and responsibility, and other outcomes relating to student development are fundamentally different goals that cannot be achieved through traditional lecture or individual study. Assessment of such outcomes is also a complex task that requires a shift in focus for instructors.

Creating assessment tools and evaluation structures that establish ways to “hear students’ voices” is important in understanding the learning experience beyond academic outcomes (ACPA and NASPA, 2004, p. 33). Students’ interpersonal skills were challenged and developed by their participation in this group project, and group members reported recognizing strengths in other classmates. In particular, feedback from English language learner students showed uniquely positive results for this population. Furthermore, research on student retention has found that “student engagement in educationally purposeful activities is positively related to academic outcomes as represented by first-year student grades and by persistence between the first and second year of college” (Kuh et al., 2008, p. 555).

Data from this investigation show that students did indeed demonstrate behaviors consistent with the four SLOs and SDOs through their participation in the capstone video project, and researchers concluded that the assignment is therefore worthy of keeping within the course curriculum. More broadly, this evaluation uncovered a depth and nuance of group dynamics frequently not emphasized in science assignments. Analysis of the interactions among group members involved in the capstone video project showed students engaged in a creative and complex process to combine diverse ideas into a single product.

Without the deliberate implementation of strategies to create a sense of community and belonging, freshman students enrolled in a large class at a large university can easily get a sense of insignificance and may discount their contribution to group endeavors. The efforts described in this study combined a focus on content learning with personal development and are important, because higher education institutions must pay attention to how students’ individual growth relates to the larger context of group behaviors. Setting specific expectations that connect individual outcomes to cooperative experiences is challenging, but in keeping with the changing demands on colleges and universities to prepare students civically as well as academically. While results here are indeed positive, it is necessary to note that the project was not the only student-centered assignment of the semester; instructors made deliberate efforts to facilitate cooperative learning across the curriculum. Specifically, while the final group video project was the largest in scale, students were required to work in groups during most class sessions, and instructors coached students on the behaviors required for successful group performance, such as listening to one another and developing roles within groups. Daily group assignments included such activities as reading discussions, interactive review of exam questions, and peer review of student writing samples.

Many research studies, however, have documented the resistance of STEM faculty to give up instructor-centered instructional strategies in favor of a more student-centered approach (Knight and Wood, 2005). Reasons for this resistance are many, but include frustrations with the demands of managing group learning and lack of evidence of its effectiveness. To replicate the positive results shown here, other instructors must be willing to adapt classroom expectations and pedagogical approaches. Cooperative group projects need not take
extensive time, elaborate equipment, or even team-taught
courses, such as the FYI course used here. It is highly recom-
mended that instructors who are new to cooperative learning
begin with short, simple projects, such as cooperative quizzes
(Jensen et al., 2002), and then progress to group projects, such
as cooperative lab reports and short group presentations from
either lecture or lab. Even informal group learning projects,
such as think–pair–share, can allow students to begin to foster
skills germane to development learning outcomes. Thought-
ful consideration of how students are assigned to lab groups
in introductory biology classes provides an opportunity for
science instructors to take advantage of a preexisting class
structure and increase the emphasis on both development
and content learning outcomes. Particularly for students en-
rolled in science classes as nonmajors, authentic learning ac-
tivities help address institutional challenges of creating “a
scientifically literate population” (McPhearson et al., 2008, p.
150) despite great variation in previous exposure to scientific
thinking and concepts and limited requirements to explore
these topics in college.

Allen and Tanner (2005) also document the resistance of
many students to increased responsibility when engaged in
active-learning strategies. The varied experiences of first-year
undergraduate students in their high school science classes
may also impact how prepared students are to work on group
projects. Given this resistance and the wide possible vari-
ety of group learning experiences, it is again important to
emphasize that the capstone assignment was not the only
group activity of the course. Much smaller-scale and lower-
stakes group work was used throughout the course. Students
learned to take responsibility for their own learning, while
also taking on the responsibility of participating in groups
in a positive manner. An awareness of personal learning and
interpersonal communication style prior to beginning group
work may aid students in their initial interactions, but scaf-
dolding opportunities in class for students to explore how to
work with others in new ways can provide a stepping-stone
toward success in higher-stakes experiences. Despite posi-
tive results in this study, an important issue not addressed is
related to the transferability of the behaviors exhibited, and
hopefully “learned,” in this one course during the completion
of one large project. Pursuing a research question related to
the transferability of behaviors germane to SDOs is a logical
next step in this expanding field of research, and consistent
with the more heavily investigated area of the transferability
of cognitive skills related to content learning objectives. In
keeping with constructivist philosophy, however, it is logical
to assume that integration of both developmental and content
outcomes into multiple parts of the curriculum will allow for
repeated opportunities for learning, and, hopefully, retention
of this learning.

Changes in what students are expected to be able to do
upon graduation, in addition to what they are expected to
know are clear. Increasing numbers of graduate and profes-
sional schools (including medical schools and programs in
other health-related career fields) require more from their
applicants than evidence of raw content knowledge—rather,
they are looking for multiple skill sets that include the abili-
ty to work on diverse teams, demonstrated appreciation of
differences, and other characteristics related to development
outcomes. The implications for instructors also clearly point
to the need for a shift in pedagogical approach. While work-
ing in groups, students used skills that are recognized as
essential for future productive citizens, scientists, and engi-
neers (see NRC, 1999, 2003b). Individual and focus group
interviews allowed students to further reflect on the group
video assignment. Many participants showed signs of per-
sonal growth relative to these indicators, and also showed
genuine enthusiasm for the project.

Undergraduate science professors should respond by mod-
ifying individualistic assignments and traditional content-
focused exams, and move to the underutilized world of co-
operative group learning. It is also through such experiences
that students engage in more authentic, project-based sci-
ence work. Scientists and medical professionals do not work
alone—they work in teams and their performance is interde-
pendent. A clear need exists for future efforts to focus more
attention on the development of effective ways to evaluate
progress toward these outcomes, to define ways to estab-
lish students’ incoming level of familiarity with cooperative
group work, and to build university-wide support for such
projects. Through participation in a final group project, stu-
dents in this study were indeed engaged in tasks that promote
progress toward both learning and development outcomes.
Other instructors are encouraged to adapt and share similar
ideas and approaches toward the goal of improving learning
for all students.

ACKNOWLEDGMENTS

This project was funded in part by a grant from the UMN’s Academy
for Distinguished Teachers. We thank Dr. Rhiannon Williams for her
assistance in evaluation design.

REFERENCES

Allen D, Tanner K (2005). Infusing active learning into the large-
enrollment biology class: seven strategies, from the simple to com-

American College Personnel Association and National Associa-
tion of Student Personnel Administrators (2004). Learning Reconsidered:
A Campus-Wide Focus on the Student Experience, Washington, DC: ACPS
and NASPA.

Baxter Magolda MB (2000). Creating Contexts for Learning and Self-
Authorship: Constructive-Developmental Pedagogy, Nashville, TN:
Vanderbilt University Press.

Student Dev 49, 269–284.

Council for Opportunity in Education. www.coenet.us/
ecm/AM/Template.cfm?Section=What_is_TRIO&Template=/CM/
HTMLDisplay.cfm&ContentID=10417 (accessed 26 January 2012).

Cresswell JW (2009). Research Design: Qualitative, Quantitative, and

DeBurman SK (2002). Learning how scientists work: experiential re-
search projects to promote cell biology learning and scientific process

literature on teaching engineering design through project-oriented

Eisen A (1998). Small-group presentations: teaching “science think-
ing” and context in a large biology class. BioScience 48, 53–58.

facilitates student retention and contributes to student learning. J
Coll Student Retention 5, 325–343.


CBE—Life Sciences Education (LSE) is an online, quarterly journal owned and published by the American Society for Cell Biology (ASCB). The journal publishes original, previously unpublished, peer-reviewed articles on life sciences education research and evidence-based practice at the K–12, undergraduate, and graduate levels. The editorial board of LSE is particularly interested in students’ conceptions and cognition related to biology and scientific inquiry and students’ engagement in the study of life sciences, as well as the efficacy of particular approaches to teaching in the life sciences. In addition, the ASCB believes that biology learning encompasses diverse fields, including math, chemistry, physics, engineering, and computer science, as well as the interdisciplinary intersections of biology with these fields. One goal of the journal is to encourage teachers and instructors to view teaching and learning the way scientists view their research, as an intellectual undertaking that is informed by systematic collection, analysis, and interpretation of data related to student learning. Target audiences include those involved in education in K–12 schools, two-year colleges, four-year colleges, science centers and museums, universities, and professional schools, including graduate students and postdoctoral researchers.

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.

All published articles are available freely online without subscription. In addition, published articles are indexed in PubMed and available through PubMed Central. A goal of LSE is to stimulate dialogue. LSE invites readers to send comments that will be considered for publication in the Letters to the Editor column.

Determining the Suitability of a Manuscript for LSE

Articles. LSE publishes articles that focus on questions of science education research and practice. Articles can either break new ground in understanding teaching and learning or describe the implementation and evaluation of educational innovations. Articles must include collection, analysis, and interpretation of 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. When possible, articles should include any newly described instructional resources and assessment instruments so that they are available for evaluation by reviewers and use by readers. The online nature of the journal facilitates the inclusion of instructional materials such as syllabi, assignments, rubrics, laboratory protocols, or professional development guidelines.

Articles detailing educational innovations must include assessment appropriate for the innovation as well as the assessment instruments, protocols, rubrics, and methods, or references to these components of the work. LSE is interested in educational innovations that have not been described extensively elsewhere, although articles about the outcomes of adopting or adapting particular approaches to teaching and learning with diverse learners and in diverse learning contexts are welcomed. Carefully constructed descriptions of relevant literature are important for demonstrating how a particular innovation is unique compared with previously published work. Instructor interested in publishing their educational innovations in LSE should give careful thought to how they will assess student learning. 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.

Because the concept of assessment may be unfamiliar to many instructors, some suggestions for assessment are presented here. Answering three questions can help guide the process of assessment: 1) What are your instructional goals? 2) What should your students know or be able to do if you met your goals? 3) How can you measure or otherwise document whether your students know or are able to do what you intend? Documentation of intended outcomes can be accomplished through systematic analysis of data collected through diverse approaches, such as pretest/posttest, interviews, focus groups, surveys, or performance on coursework, including exams, papers, or lab reports. Authors should present their innovations in the same way that life scientists present their research: claims regarding efficacy should be supported by evidence.

Answering the following questions can help you determine whether your manuscript is ready for consideration by LSE: What is the background or motivation for your project? Does it meet a previously unmet need or provide a test of some educational theory? What was your goal in developing it? How do you know if you have succeeded? Can you show success to readers? Are the outcomes and supporting evidence sufficient to convince others to adapt your materials, approach, or project for their own use? Are the outcomes sufficiently novel to warrant publication?

Essays. LSE publishes essays, which describe approaches to educational challenges, 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.

Meeting Reports. The journal publishes occasional reports of meetings related to education in the life sciences. 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.

Letters to the Editor. LSE welcomes letters as a means for initiating discussion of issues of interest to readers. LSE requires...
that letters are directly responsive to an article published in the journal or that they bring a new or under-recognized issue to the attention of readers in a way that is informed by the inclusion of relevant data or references to the literature. Letters received through the normal online submission process will be reviewed by the Editor-in-Chief and published at her discretion. Appropriate topics for letters include commentary on education-related articles in this or other journals as well as opinions on more general topics of concern to the readership.

Features. Articles listed under the heading of “Features” are by invitation only. Authors interested in contributing to a feature should contact the editor.

Announcements. LSE accepts noncommercial announcements of meetings, workshops, and conferences and of funding opportunities and fellowships open to all.

Peer-Review Process

All submitted manuscripts and educational materials are subject to peer review. After a manuscript has been submitted to LSE, the Editor-in-Chief selects an editorial board member to guide the paper through the review process. Editorial board members select two reviewers to submit written evaluations. The board member will assess the peer reviews and determine whether the submission will be accepted as is, accepted with suggested revisions, temporarily rejected with suggestions for improvements before resubmission, or rejected with reasons explaining this decision. The corresponding author can usually expect an initial response within four weeks.

License and Publishing Agreement

Authors are required to sign a License and Publishing Agreement when a manuscript is accepted for publication. Under the terms of that agreement, authors retain copyright but grant the ASCB a perpetual license to publish the manuscript. Authors also grant to the public the nonexclusive right to copy, distribute, or display the manuscript subject to the terms of the Creative Commons–Noncommercial–Share Alike 3.0 Unported license (http://creativecommons.org/licenses/by-nc-sa/3.0).

ASCB Policy on Research Misconduct by Authors

By submitting a paper to LSE, an author acknowledges that he or she is subject to the ASCB Policy on Research Misconduct by Authors. The policy is posted at www.ascb.org/files/research_misconduct.pdf.

Institutional Review Board Approval

Manuscripts reporting on studies involving human subjects must include explicit assurance that the research reported was approved by a local Institutional Review Board, including an Institutional Review Board number, unless the research is exempt from such review according to U.S. Department of Education guidelines. Prospective authors are advised that permission must be obtained in advance.

Guidelines for Preparing Articles, Essays, and Features

General Instructions. Submitted papers must be original (i.e., not published or submitted for publication elsewhere). Previously published material will not be considered. Publication of a short abstract does not constitute prior publication, nor does presentation of data at a scientific meeting or in a webcast of such a meeting. The posting of a manuscript on an author’s personal website or in an online institutional repository prior to submission of the manuscript to LSE is not considered prior publication. Publication of a paper in the proceedings of a scientific meeting generally does constitute prior publication.

The text should be written in clear, concise, and grammatical English. Manuscripts ordinarily begin with an overview of how the work presented is relevant to the classroom, laboratory, or curriculum and what student outcomes are expected. Whenever possible, incorporate materials by citing relevant publications, without repeating already published works. Manuscript files must be submitted in either .doc or .rtf format.

Tables and illustrations should convey information effectively. Graphs and figures should be provided digitally as separate TIF or EPS files. Authors are encouraged to take advantage of the online nature of LSE. Video, audio, databases, images, animations, molecular structures, and other electronic resources may be linked as supplemental material for further consideration by readers.

International authors may wish to consider using an editorial service, such as ScienceDocs (www.sciencedocs.com), The Medical Editor (www.themedicaleditor.com), American Journal Experts (www.journalexperts.com), Editage (www.editage.com), Bioscience Writers (www.biosciencewriters.com), 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 reviewers (please include institution and email address).

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), the name of the corresponding...
author, and the names and affiliations of all other authors. For the corresponding author, give complete postal and email addresses 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 student group (primary, secondary, undergraduate, graduate, or postdoc) and a second keyword that indicates which cells or organism are studied in the course or lab exercise (e.g., bacteria, fungi, plants, yeast, Drosophila, nematodes).

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. All pages must be numbered.

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.

Citations to online websites are acceptable reference links. In the text, include the name of author for the cited website and the date accessed (e.g., National Center for Biotechnology Information, 2005). The citation should include the author (if known), date of latest modification (if known) or copyright, title of page, title of complete work (if appropriate), URL (e.g., www.website.org), and date of access. For example: National Center for Biotechnology Information (2005). NCBI Home Page. www.ncbi.nlm.nih.gov (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’ laboratory. If you cite no references, include the results of a literature search that cites your own and others’ work that has a bearing on the manuscript you are submitting and explain why no citations are needed. Include in this section only material that has been published in the literature or on the Internet.

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. Table files need to be uploaded separately.

Supplemental Material. Upload all supplemental material (except in the case of videos) together in one combined PDF.

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. Assessment may come in the form of learning gains (e.g., pre- and posttests) and/or changes in student attitudes toward the electronic work. Length of time spent using the electronic work may be one indicator of student attitudes. If it is possible to measure student responses to online questions, these responses (number of submissions before correct answer, typical errors, etc.) may further a reader’s understanding of how the electronic work could be incorporated into his or her teaching.

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 manuscript should include an abstract and any figures deemed helpful when the published work appears as a PDF file.

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, and monoclonal antibodies 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. Instructional materials that were developed in the course of a project described in a LSE article should also be made freely available, preferably in electronic format. 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 either .doc 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.
A searchable collection of images, videos, and animations of cells and cellular functions available to students, educators, and the interested public.

New Features with Optional Free Accounts
- Save favorite images for later use
- Automatically find new images that match your areas of interest

Coming Soon!
- Annotate images
- Invite colleagues to connect and share images
- Create multiple folders for saving/sharing images
- Subscribe to receive emails when new matching images are available

Submit your images and feedback and visit www.cellimagelibrary.org for more information. Or email David Orlaff (dorlaff@ncmi.ucsd.edu) at the Image Library.

Questions for Stimulating Inquiry*

There are many types of cells, and although they have numerous similarities, they also display profound differences in shape, structure, and function. These images highlight contrasts and commonalities across a spectrum of organisms. Use them as a springboard for deeper exploration of cells and biological processes.

1. Which panels display cells that are differentiated, that is, specialized for particular functions? How many of their functions can you name? Would you argue that the amoeba is a differentiated cell, or not?

2. Which panels display cells that are not differentiated? As they continue to divide, will some of their descendants become differentiated? What kinds of differentiated cells will they be?

3. How many different kinds of organelles can you identify in these images? What are their functions?

4. How do the forms of the cells in these images relate to their functions? For example, why do the two different types of blood cells have different shapes? How about the others?

5. Using the cell diameters given in the captions, how can you determine the final magnification of each image?

*For answers and discussion, see the CBE-LSE website: www.lifescied.org

All images taken by Steven S. Hsu, Department of Biology, California State University, Fresno. bw micrograph: Marchantia polymorpha, pigmented alga. Arrow indicates chloroplast. Top arrow indicates a cell containing a spore. © 2003 Stanford University. Amoeba proteus stained with pH-dependent dye (0.7 mm): micrograph taken by Brian E. Cowles. © 2007 Stanford University. Sea urchin embryo (Lytechinus pictus) in second cell division (0.1 mm): micrograph taken by Brian E. Cowles. © 2007 Stanford University. Human cheek cells stained with methylene blue (0.03 mm): micrograph taken by Amy J. Tilley. © 2008 Stanford University. Human red blood cells (0.008 mm): micrograph taken by 1941-2007, Kamiya Research Institute. Microscopy World: Prentice Hall. Xenopus laevis, 1.3 mm: micrograph taken by Stan Marchant. Xenopus laevis, 0.015 mm: micrograph taken by Stan Marchant. Xenopus laevis, 0.025 mm: micrograph taken by Stan Marchant. © 2007 Stanford University. Early zebrafish embryo (Danio rerio), 2-cell stage, shell removed (0.5 mm): micrograph taken by Brian E. Cowles. © 2007 Stanford University. Human blood cells (red blood cells and granulocytes, latter 0.015 mm): micrograph taken by 1991-2007, Kamiya Research Institute. Microscopy World: Prentice Hall.
MBoC
MOLECULAR BIOLOGY OF THE CELL

Brought to you by the American Society for Cell Biology and run entirely by scientists for scientists

IMPORTANT
MBoC presents conceptual advances of broad interest within all areas of cell biology.

INTERNATIONAL
The Editorial Board includes Associate Editors in twelve countries.

ACCESSIBLE
Each issue is freely available two months after publication. Authors’ manuscripts are posted within weeks of acceptance.

AUTHOR FRIENDLY
MBoC offers fair, rapid, constructive peer review; free color figures; and a 20% discount on page charges for ASCB members.

NOW WITH EXCITING NEW FEATURES:
- Retrospectives and Perspectives
- MBoC Technical Perspectives
- ASCB Award Essays
- ASCB Annual Meeting Highlights
- And much more

Visit Our Redesigned Website!

www.molbiolcell.org
The HHMI Exceptional Research Opportunities Program aims to increase the diversity of the science professoriate. We match well-prepared undergraduate students with HHMI scientists in a summer-long mentored research experience. Since 2003, EXROP has sponsored 578 students from 122 colleges and universities who worked with 187 HHMI scientists. More than three-fourths of the students are underrepresented minorities. Of the 370 EXROP alumni who have completed the baccalaureate degree, 362 are in science careers including 186 who went into graduate programs.
<table>
<thead>
<tr>
<th>Name</th>
<th>Institution 1</th>
<th>Institution 2</th>
<th>Institution 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carlos Hernández</td>
<td>Columbia University</td>
<td>J. Andrew McCorman</td>
<td>University of California, San Diego</td>
</tr>
<tr>
<td>Meryl Horwitz</td>
<td>City University of New York Hunter College</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Li-Huei Tsai</td>
<td>Massachusetts Institute of Technology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dalton Hughes</td>
<td>University of Maryland Baltimore County</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Isiah Warner</td>
<td>Louisiana State University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Victoria Jones</td>
<td>North Carolina Central University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oliver Hobert</td>
<td>Columbia University College of Physicians and Surgeons</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kristin Kendrick</td>
<td>Louisiana State University and A &amp; M College</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cornelia I. Bargmann</td>
<td>The Rockefeller University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jennifer Ky</td>
<td>Yale University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Michael B. Elowitz</td>
<td>California Institute of Technology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diana Lenis</td>
<td>Stony Brook University</td>
<td>Joan A. Steitz</td>
<td>Yale School of Medicine</td>
</tr>
<tr>
<td>Lanair Lett</td>
<td>Harvard University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Daniel A. Melton</td>
<td>Harvard University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kimberly Loo</td>
<td>University of California-Los Angeles</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matthew L. Warman</td>
<td>Children’s Hospital, Boston</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Melany Lopez</td>
<td>Washington University in St. Louis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>David E. Clepham</td>
<td>Children’s Hospital, Boston</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mariana Lopez-Rosas</td>
<td>Yale University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eric R. Kandel</td>
<td>Columbia University College of Physicians and Surgeons</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kristine Lynn</td>
<td>Lewis and Clark College</td>
<td>David D. Ginty</td>
<td>The Johns Hopkins University School of Medicine</td>
</tr>
<tr>
<td>Emmanuel Mendoza</td>
<td>UC Davis</td>
<td>Leonard L. Zon</td>
<td>Children’s Hospital, Boston</td>
</tr>
<tr>
<td>Cristina Murphy</td>
<td>Louisiana State University</td>
<td>Richard Axel</td>
<td>Columbia University College of Physicians and Surgeons</td>
</tr>
<tr>
<td>Brittni Naylor</td>
<td>Louisiana State University</td>
<td>Christopher Y. Plowe</td>
<td>University of Maryland School of Medicine</td>
</tr>
<tr>
<td>Anthony Nguyen</td>
<td>Brandeis University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Joachim Frank</td>
<td>Columbia University College of Physicians and Surgeons</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sang Nguyen</td>
<td>University of California-Riverside</td>
<td>George Q. Daley</td>
<td>Children’s Hospital, Boston</td>
</tr>
<tr>
<td>Maria Nguyen</td>
<td>Oregon State University</td>
<td>Joseph G. Gleeson</td>
<td>UC San Diego School of Medicine</td>
</tr>
<tr>
<td>Kaitlyn Norman</td>
<td>New Mexico State University</td>
<td>James J. Collins</td>
<td>Boston University</td>
</tr>
<tr>
<td>Tinashe Nyamhete</td>
<td>Colgate University</td>
<td>Matthew P. Scott</td>
<td>Stanford University School of Medicine</td>
</tr>
<tr>
<td>Fiona Odu</td>
<td>University of Missouri Columbia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Luis A. Nunes Amoral</td>
<td>Northwestern University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alejandra Olvera</td>
<td>Wesleyan University</td>
<td>Pamela J. Birkman</td>
<td>California Institute of Technology</td>
</tr>
<tr>
<td>Joshua Onyango</td>
<td>Oakwood University</td>
<td>Linda B. Buck</td>
<td>Fred Hutchinson Cancer Research Center</td>
</tr>
<tr>
<td>Fabian Ortega</td>
<td>Yale University</td>
<td>David G. Schatz</td>
<td>Yale School of Medicine</td>
</tr>
<tr>
<td>Jennifer Ortega</td>
<td>Haverford College</td>
<td>Louis J. Ptacek</td>
<td>UC San Francisco</td>
</tr>
<tr>
<td>Justin Paz</td>
<td>University of Minnesota-Twin Cities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carlos Peinado</td>
<td>UC San Diego</td>
<td>Susan M. Koech</td>
<td>Yale University</td>
</tr>
<tr>
<td>Fernando Pimentel</td>
<td>Stanford University</td>
<td>Stephen R. Goff</td>
<td>Columbia University College of Physicians and Surgeons</td>
</tr>
<tr>
<td>Michael Polen</td>
<td>Widener University</td>
<td>Yi Lu</td>
<td>University of Illinois at Urbana-Champaign</td>
</tr>
<tr>
<td>Amanda Rabe</td>
<td>University of California-Santa Cruz</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Samuel Regalado</td>
<td>UC Berkeley</td>
<td>Amy J. Wagers</td>
<td>Harvard University</td>
</tr>
<tr>
<td>Donald Richards</td>
<td>UC Riverside</td>
<td>Leslie A. Leinwand</td>
<td>University of Colorado at Boulder</td>
</tr>
<tr>
<td>Jesse Robison</td>
<td>Virginia Polytechnic Institute &amp; State University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Luis A. Nunes Amoral</td>
<td>Northwestern University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lauren Rodriguez</td>
<td>UC Santa Cruz</td>
<td>William R. Jacobs Jr.</td>
<td>Albert Einstein College of Medicine</td>
</tr>
<tr>
<td>Juan Ruiz</td>
<td>University of Miami</td>
<td>Jeffrey Karp</td>
<td>Brigham and Women’s Hospital</td>
</tr>
<tr>
<td>Sonja Sabater</td>
<td>Duke University</td>
<td>Kerry J. Ressler</td>
<td>Emory University School of Medicine</td>
</tr>
<tr>
<td>Elisha Sanders</td>
<td>University of Washington</td>
<td>Hugo J. Bellen</td>
<td>Baylor College of Medicine</td>
</tr>
<tr>
<td>Amy Schade</td>
<td>University of North Texas</td>
<td>Paul Ahiquist</td>
<td>University of Wisconsin-Madison</td>
</tr>
<tr>
<td>Victoria Senechal</td>
<td>University of California, Riverside</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Richard L. Huganir</td>
<td>Johns Hopkins University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swastina Shrestha</td>
<td>Mount Holyoke College</td>
<td>Joanna Chory</td>
<td>Salk Institute for Biological Studies</td>
</tr>
<tr>
<td>Angelica Simmons</td>
<td>Louisiana State University</td>
<td>Vivian Cheung</td>
<td>Children’s Hospital, Philadelphia</td>
</tr>
<tr>
<td>Julia Smith</td>
<td>Georgetown College</td>
<td>Erich D. Jarvis</td>
<td>Duke University</td>
</tr>
<tr>
<td>Projita Tamir</td>
<td>College of William and Mary</td>
<td>Randall T. Moon</td>
<td>University of Washington School of Medicine</td>
</tr>
<tr>
<td>Sofie Tucker</td>
<td>Yale University</td>
<td>Shawn Ferguson</td>
<td>Yale School of Medicine</td>
</tr>
<tr>
<td>Theodor Uzamere</td>
<td>Morehouse College</td>
<td>Tyler Jacobs</td>
<td>Massachusetts Institute of Technology</td>
</tr>
<tr>
<td>Alana Van Dervort</td>
<td>Florida International University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Douglas A. Melton</td>
<td>Harvard University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shu Wang</td>
<td>Wesleyan University</td>
<td>Martin D. Burke</td>
<td>University of Illinois at Urbana-Champaign</td>
</tr>
<tr>
<td>Cherise Washington</td>
<td>Spelman College</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Charles S. Zuck</td>
<td>Columbia University College of Physicians and Surgeons</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monique Williams</td>
<td>Oakwood University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alejandro Sánchez Alvarado</td>
<td>Stowers Institute for Medical Research</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*The HHMI EXROPS Capstone Award provides funding for EXROPS students to complete a second summer research experience with their EXROPS mentor.*

Email: exrop@hhmi.org
http://www.hhmi.org
ESSAY
Writing-to-Learn in Undergraduate Science Education: A Community-Based, Conceptually Driven Approach
Julie A. Reynolds, Christopher Thaiss, Wendy Katkin, and Robert J. Thompson, Jr. ............................................... 41–49

ARTICLES
Biology Undergraduates’ Misconceptions about Genetic Drift

The C.R.E.A.T.E. Approach to Primary Literature Shifts Undergraduates’ Self-Assessed Ability to Read and Analyze Journal Articles, Attitudes about Science, and Epistemological Beliefs
Sally G. Hoskins, David Lopatto, and Leslie M. Stevens .................................................................................. 62–72

Exploring Undergraduates’ Understanding of Photosynthesis Using Diagnostic Question Clusters
Joyce M. Parker, Charles W. Anderson, Merle Heidemann, John Merrill, Brett Merritt, Gail Richmond, and Mark Urban-Lurain .......................................................... 73–83

Applying Computerized-Scoring Models of Written Biological Explanations across Courses and Colleges: Prospects and Limitations
Minsu Ha, Ross H. Nehm, Mark Urban-Lurain, and John E. Merrill ................................................................ 84–98

Using Comparative Genomics for Inquiry-Based Learning to Dissect Virulence of Escherichia coli O157:H7 and Yersinia pestis

Multiple-Choice Exams: An Obstacle for Higher-Level Thinking in Introductory Science Classes
Kathrin F. Stanger-Hall .......................................................................................................................... 112–124

Does Instructor Type Matter? Undergraduate Student Perception of Graduate Teaching Assistants and Professors
K. Denise Kendall and Elisabeth E. Schussler ............................................................................................... 125–137

Using Student Learning and Development Outcomes to Evaluate a First-Year Undergraduate Group Video Project
Murray Jensen, Allison Mattheis, and Brady Johnson .................................................................................. 138–150