# HIGHLIGHTS OF 2014

## Table of Contents

### Editorial
- Thanks!
  - Erin L. Dolan ................................................................. 1–2

### FEATURES
- **Letter to the Editor**
  - Interactions Are Critical
    - Christopher W. Beck and Nancy G. Bliwise ............................................................... 3–4

- **Approaches to Biology Teaching and Learning**
  - Teaching More by Grading Less (or Differently)
    - Jeffrey Schinske and Kimberly Tanner ................................................................. 5–12

- **WWW. Life Sciences Education**
  - Engaging with Molecular Form to Understand Function
    - Nicola C. Barber and Louisa A. Stark ................................................................. 13–16

- **From the National Science Foundation**
  - Vision and Change in the Biology Community: Snapshots of Change
    - Helen L. Vasaly, Jason Feser, Matthew D. Lettrich, Kevin Correa, and Katherine J. Denniston .......... 17–21

- **Meeting Report**
  - Assessment of Course-Based Undergraduate Research Experiences: A Meeting Report

### RESEARCH METHODS
- Is It the Intervention or the Students? Using Linear Regression to Control for Student Characteristics in Undergraduate STEM Education Research
  - Roddy Theobald and Scott Freeman ................................................................. 34–41

### ESSAY
- Feedback about Teaching in Higher Ed: Neglected Opportunities to Promote Change
  - Cara Gormally, Mara Evans, and Peggy Brickman ................................................................. 42–54

*Table of Contents Continued on Back Cover*
NEW SIXTH EDITION

MOLECULAR BIOLOGY OF THE CELL

BRUCE ALBERTS, University of California, San Francisco, USA
ALEXANDER JOHNSON, University of California, San Francisco, USA
JULIAN LEWIS, Formerly of Cancer Research, UK
DAVID MORGAN, University of California, San Francisco, USA
MARTIN RAFF, University College London, UK
KEITH ROBERTS, Emeritus, University of East Anglia, UK
PETER WALTER, University of California, San Francisco, USA

As the amount of information in biology expands dramatically, it becomes increasingly important for textbooks to distill the vast amount of scientific knowledge into concise principles and enduring concepts. As with previous editions, Molecular Biology of the Cell, Sixth Edition accomplishes this goal with clear writing and beautiful illustrations. The Sixth Edition has been extensively revised and updated with the latest research in the field of cell biology, and it provides an exceptional framework for teaching and learning.

For more information and to purchase at a 30% DISCOUNT, visit www.garlandscience.com/mboc6 and apply PROMO CODE AGL91 at checkout.

Or visit us at BOOTH #909 at the ASCB meeting.

CONTENTS

INTRODUCTION TO THE CELL
1. Cells and Genomes
2. Cell Chemistry and Bioenergetics
3. Proteins
BASIC GENETIC MECHANISMS
4. DNA, Chromosomes, and Genomes
5. DNA Replication, Repair, and Recombination
6. How Cells Read the Genome: From DNA to Protein
7. Control of Gene Expression
WAYS OF WORKING WITH CELLS
8. Analyzing Cells, Molecules, and Systems
9. Visualizing Cells
INTERNAL ORGANIZATION OF THE CELL
10. Membrane Structure
11. Membrane Transport of Small Molecules and the Electrical Properties of Membranes
12. Intracellular Compartments and Protein Sorting
13. Intracellular Membrane Traffic
14. Energy Conversion: Mitochondria and Chloroplasts
15. Cell Signaling
16. The Cytoskeleton
17. The Cell Cycle
18. Cell Death
CELLS IN THEIR SOCIAL CONTEXT
19. Cell Junctions and the Extracellular Matrix
20. Cancer
21. Development of Multicellular Organisms
22. Stem Cells and Tissue Renewal
23. Pathogens and Infection
24. The Innate and Adaptive Immune Systems

www.garlandscience.com
CBE—Life Sciences Education (ISSN 1931-7913) is published online quarterly (one volume per year) by The American Society for Cell Biology, 8120 Woodmont Avenue, Suite 750, Bethesda, MD 20814-2762, USA; www.ascb.org. The journal, which is accessible at www.lifescied.org and www.pubmedcentral.nih.gov, is supported in part by a grant from the Howard Hughes Medical Institute. “ASCB,” “The American Society for Cell Biology,” “iBioSeminars,” and “Molecular Biology of the Cell” are registered trademarks of The American Society for Cell Biology. “The Cell: An Image Library” is a common law trademark of The American Society for Cell Biology.

Owned and published by The American Society for Cell Biology, Stefano Bertuzzi, Executive Director
A Peer-Reviewed Journal of Biology Education Research and Evidence-Based Biology Teaching

For educators at all levels and across all life science disciplines

Online and completely free—no subscription required

“Leaders of biology education value *LSE’s* high standards and ability to reach the ‘grassroots’—people who teach biology. Our approach is both thoughtful and developmental and leads to exemplary biology education studies.”

—Erin Dolan, Editor-in-Chief, Executive Director, Texas Center for Science Discovery, UT-Austin

Publish for free in *LSE*

*LSE* emphasizes teaching innovations and evidence of their effectiveness. It publishes original research articles, essays, and features that help you apply education research to your own teaching.

*LSE* is supported in part by a grant from the Howard Hughes Medical Institute.
Editorial

Thanks!

Erin L. Dolan

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

In my 2013 editorial, I highlighted several “firsts” experienced by CBE—Life Sciences Education (LSE) and the broader biology education community. I then challenged readers to embark on some additional firsts aimed at encouraging others to make use of evidence-based instruction. This year, I would like to thank everyone for embracing this challenge and continuing to raise the visibility of effective ways to teach and learn biology. All indications are that your efforts, coupled with investment of the American Society for Cell Biology (ASCB), the Howard Hughes Medical Institute, and the Genetics Society of America (GSA), are paying off: LSE is becoming more widely read and recognized.

Thanks to our authors—a group that continues to grow. Analysis of author data from 2006 to 2013 (Figure 1) shows that LSE consistently publishes work from more than 150 authors per year, with greater than 60% being first-time contributors. LSE authors represent a wide range of institution types, including both U.S. and international institutions (Table 1).

Thanks also to our reviewers. The number of LSE reviewers is also growing and now represents 250+ institutions and five countries. Even with an influx of approximately 50% new reviewers each year, LSE editors and staff continue to receive positive feedback about the constructive nature of reviews, even from authors of manuscripts not selected for publication. In fact, LSE editors and staff regularly receive positive feedback from biologists indicating that the journal publishes work that is approachable to them, in contrast to journals published by and for those formally trained as science education researchers.

Thanks to ASCB and GSA leadership for ensuring that articles published in LSE are increasingly “on the radar” of all biologists. For example, ASCB members receive alerts when each quarterly issue goes live, and each LSE table of contents is published in the ASCB Newsletter. GSA is featuring one LSE article per month in the e-highlights (www.genetics.org/content/198/1/NP.full) of its journal, Genetics. The quality and groundbreaking nature of the work published in LSE...
Table 1. Author institution typesa

<table>
<thead>
<tr>
<th>Carnegie classification</th>
<th>Number (and percent) of authorship by institution type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research universities (very high research activity according to aggregated and per capita activity)</td>
<td>553 (41%)</td>
</tr>
<tr>
<td>Master's colleges and universities</td>
<td>171 (13%)</td>
</tr>
<tr>
<td>Baccalaureate</td>
<td>139 (10%)</td>
</tr>
<tr>
<td>Research universities (high research activity, high for only aggregated or per capita activity)</td>
<td>78 (6%)</td>
</tr>
<tr>
<td>International</td>
<td>73 (5%)</td>
</tr>
<tr>
<td>State or federal agency</td>
<td>53 (4%)</td>
</tr>
<tr>
<td>Doctoral universities</td>
<td>40 (3%)</td>
</tr>
<tr>
<td>Research institute</td>
<td>21 (1%)</td>
</tr>
<tr>
<td>Company</td>
<td>15 (1%)</td>
</tr>
<tr>
<td>Professional schools (health, engineering)</td>
<td>14 (1%)</td>
</tr>
<tr>
<td>Professional society</td>
<td>13 (1%)</td>
</tr>
<tr>
<td>Associates</td>
<td>12 (1%)</td>
</tr>
<tr>
<td>K–12 schools</td>
<td>12 (1%)</td>
</tr>
<tr>
<td>Museum</td>
<td>7 (0.5%)</td>
</tr>
<tr>
<td>Other (e.g., foundations, no affiliation, retired)</td>
<td>125 (9%)</td>
</tr>
<tr>
<td>Total</td>
<td>1340 authorships (100%)</td>
</tr>
</tbody>
</table>

aThese data represent all authorship between 2006 and 2013, rather than individual authors represented in Figure 1. Each author of each paper is counted such that a single author will be counted multiple times if he or she has authored multiple papers.

Even with all of this progress, teaching strategies that have been shown to be effective are still not widely used. Many faculty members are unaware that these strategies exist and have been shown to be effective (Henderson et al., 2011). Others lack knowledge of how to use this research to improve instruction (Andrews et al., 2011). Institutional and cultural barriers slow or even prevent faculty members from adopting teaching methods demonstrated to be beneficial, precluding widespread employment of effective instructional practices (Brownell and Tanner, 2012). To address this, we need to continue to advocate for evidence-based instruction, as well as publish research that breaks new ground in understanding biology teaching and learning and educational innovations that have been evaluated for efficacy. Thanks to all for your extraordinary efforts to bring everyone under the tent of evidence-based instruction.

REFERENCES


Letter to the Editor

Interactions Are Critical
Christopher W. Beck* and Nancy G. Bliwise†

*Department of Biology and †Department of Psychology, Emory University, Atlanta, GA 30322

To the Editor:

Recently, Theobald and Freeman (2014) reviewed approaches for measuring student learning gains in science, technology, engineering, and mathematics (STEM) education research. In their article, they highlighted the shortcomings of approaches such as raw change scores, normalized gain scores, normalized change scores, and effect sizes when students are not randomly assigned to classes based on the different pedagogies that are being compared. As an alternative, they propose using linear regression models in which characteristics of students, such as pretest scores, are included as independent variables in addition to treatments. Linear models that include both continuous and categorical independent variables are often termed analysis of covariance (ANCOVA) models. The approach of using ANCOVA to control for differences in students among treatments groups has been suggested previously by Weber (2009). We largely agree with Theobald and Freeman (2014) and Weber (2009) that ANCOVA models are an appropriate method for situations in which students cannot be randomly assigned to treatments and controls. However, in describing how to implement linear regression models to examine student learning gains, Theobald and Freeman (2014) ignore a fundamental assumption of ANCOVA: ANCOVA assumes homogeneity of slopes (McDonald, 2009; Sokal and Rohlf, 2011). In other words, the slope of the relationship between the covariate (e.g., pretest score) and the dependent variable (e.g., posttest score) is the same for the treatment group and the control. This assumption is a strict assumption of ANCOVA in that violations of this assumption can result in incorrect conclusions (Engqvist, 2005). For example, in Figure 1, both pretest score and treatment have statistically significant main effects in a linear model with only pretest score ($F(1, 97) = 25.6, p < 0.001$) and treatment ($F(1, 97) = 42.6, p < 0.01$) as independent variables. Therefore, we would conclude that all students in the class with pedagogical innovation had significantly greater posttest scores than those students in the control class for a given pretest score. Furthermore, we would conclude that the pedagogical innovation led to the same increase in score for all students in the treatment class, independent of their pretest scores. Clearly, neither of these conclusions would be justified.

Researchers must first test the assumption of the homogeneity of slopes by including an interaction term (covariate × treatment) in their linear model (McDonald, 2009; Weber 2009; Sokal and Rohlf, 2011). For example, if we measured student achievement in two courses with different instructional approaches in a typical pretest/posttest design, then the interaction between students’ pretest scores and the type of instruction must be considered, because the instruction may have a different effect for high- versus low-achieving students. If multiple covariates are included in the linear model (see Equation 1 in Theobald and Freeman, 2014), then interaction terms need to be included for each of the covariates in the model. If the interaction term is statistically significant, this suggests that the relationship between the covariate and the dependent variable is different for each treatment group ($F(1, 96) = 25.1, p < 0.001$; Figure 1). As a result, the effect of the treatment will depend on the value of the covariate, and universal statements about the effect of the treatment are not appropriate (Engqvist, 2005). If the interaction term is not statistically significant, it should be removed from the model and the analysis rerun without the interaction term. Failure to remove an interaction term that was not statistically significant also can lead to an incorrect conclusion (Engqvist, 2005). Whether there are statistically significant interactions between the “treatment” and the covariates in the data set used by Theobald and Freeman (2014) is unclear.

In addition to being a strict assumption of ANCOVA, testing for homogeneity of slopes in a linear model is important in STEM education research, as slopes are likely heterogeneous for several reasons. First, for many instruments used in STEM education research, high-achieving students score high on the pretest. As a result, their ability to improve is limited due to the ceiling effect, and differences between treatment and control groups in posttest scores are likely to be minimal (Figure 1). In contrast, low-achieving students have a greater opportunity to change their scores between their pretest and posttest scores.

DOI: 10.1187/cbe.14-05-0086
Address correspondence to: Christopher Beck (christopher.beck@emory.edu).

© 2014 C. W. Beck and N. G. Bliwise. CBE—Life Sciences Education © 2014 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. Simulated data to demonstrate heterogeneity of slopes. Pretest values were generated from random normal distributions with mean = 59.8 (SD = 18.1) for the treatment course and mean = 59.3 (SD = 17.0) for the control course, based on values given in Theobald and Freeman (2014). For the treatment course, posttest values were calculated using the formula posttest<sub>i</sub> = 80 + 0.1 × pre-test<sub>i</sub> + ε<sub>i</sub>, where ε<sub>i</sub> was selected from a random normal distribution with mean = 0 (SD = 10). For the control course, posttest values were calculated using the formula posttest<sub>i</sub> = 42 + 0.5 × pre-test<sub>i</sub> + ε<sub>i</sub>, where ε<sub>i</sub> was selected from a random normal distribution with mean = 0 (SD = 10). n = 50 for both courses.


REFERENCES


Engqvist L (2005). The mistreatment of covariate interaction terms in linear model analyses of behavioural and evolutionary ecology studies. Anim Behav 70, 967–971.


Feature
Approaches to Biology Teaching and Learning

Teaching More by Grading Less (or Differently)
Jeffrey Schinske* and Kimberly Tanner†

*Department of Biology, De Anza College, Cupertino, CA 95014; †Department of Biology, San Francisco State University, San Francisco, CA 94132

INTRODUCTION
When we consider the practically universal use in all educational institutions of a system of marks, whether numbers or letters, to indicate scholastic attainment of the pupils or students in these institutions, and when we remember how very great stress is laid by teachers and pupils alike upon these marks as real measures or indicators of attainment, we can but be astonished at the blind faith that has been felt in the reliability of the marking systems.

—I. E. Finkelstein (1913)

If your current professional position involves teaching in a formal classroom setting, you are likely familiar with the process of assigning final course grades. Last time you assigned grades, did you assign an “E,” “E+,” or “E−” to any of your students? Likely you assigned variations on “A’s,” “B’s,” “C’s,” “D’s,” and “F’s.” Have you wondered what happened to the “E’s” or talked with colleagues about their mysterious absence from the grading lexicon? While we often commiserate about the process of assigning grades, which may be as stressful for instructors as for students, the lack of conversation among instructors about the mysterious omission of the “E” is but one indicator of the many tacit assumptions we all make about the processes of grading in higher education.

Given that the time and stress associated with grading has the potential to distract instructors from other, more meaningful aspects of teaching and learning, it is perhaps time to begin scrutinizing our tacit assumptions surrounding grading. Below, we explore a brief history of grading in higher education in the United States. This is followed by considerations of the potential purposes of grading and insights from research literature that has explored the influence of grading on teaching and learning. In particular, does grading provide feedback for students that can promote learning? How might grades motivate struggling students? What are the origins of norm-referenced grading—also known as curving? And, finally, to what extent does grading provide reliable information about student learning and mastery of concepts? We end by offering four potential adjustments to our general approach to grading in undergraduate science courses for instructors to consider.

A BRIEF HISTORY OF GRADING IN HIGHER EDUCATION

It can be easy to perceive grades as both fixed and inevitable—without origin or evolution … Yet grades have not always been a part of education in the United States.

—Schneider and Hutt (2013)

Surprisingly, the letter grades most of us take for granted did not gain widespread popularity until the 1940s. Even as late as 1971, only 67% of primary and secondary schools in the United States used letter grades (National Education Association, 1971). It is therefore helpful to contextualize the subject to appreciate the relatively young and constantly changing nature of current systems of grading. While not an exhaustive history, the sections below describe some of the main developments leading to the current dominant grading system.

Early 19th Century and Before
The earliest forms of grading consisted of exit exams before awarding of a degree, as seen at Harvard as early as 1646 (Smallwood, 1935). Some schools also awarded medals based on competitions among students or held regular competitions to assign seats in class (Cureton, 1971). Given that universities like Yale and Harvard conducted examinations and elected valedictorians and salutatorians early in the 18th century, some scale of grading must have existed. However, the first official record of a grading system surfaces in 1785 at Yale, where seniors were graded into four categories: Optimii, second Optimii, Inferiores, and Perjores (Stiles, 1901, cited by Smallwood, 1935). By 1837, Yale was also recording grades...
student credit for individual classes, not just at the completion of college studies, using a four-point scale. However, these “merit marks” were written in code and hidden from students (Bagg, 1871).

Harvard and other schools soon experimented with public rankings and evaluations, noting that this resulted in “increasing [student] attention to the course of studies” and encouraged “good moral conduct” (Harvard University, 1832). Concerned that such public notices would inspire competition among students, which would distract from learning, other schools used more frequent, lower-stakes “report cards” to provide feedback on achievement (Schneider and Hutt, 2013). In 1837, at least some professors at Harvard were grading using a 100-point system (Smallwood, 1935). During this same period, William and Mary placed students in categories based on attendance and conduct. The University of Michigan experimented with a variety of grading systems in the 1850s and 1860s, including various numeric and pass/fail systems (Smallwood, 1935). Still, many schools at this time kept no formal records of grades (Schneider and Hutt, 2013).

Late 19th Century and 20th Century

With schools growing rapidly in size and number and coordination between schools becoming more important, grades became one of the primary means of communication between institutions (Schneider and Hutt, 2013). This meant grades needed to have meaning not just within an institution but also to distant third parties. A record from 1883 indicates a student at Harvard received a “B,” and in 1884, Mount Holyoke was grading on a system including “A,” “B,” “C,” “D,” and “E.” Each letter corresponded to a range of percentage scores, with lower than 75% equating to an “E” and indicating failure. Mount Holyoke added an “F” grade (for failing) to the scale in 1898 and adjusted the percentages relating to the other letters (Smallwood, 1935). This appears to be the initial origin of the “A”–“F” system familiar to most faculty members today, albeit including an “E” grade. By 1890, the “A”–“E” system had spread to Harvard after faculty members expressed concerns regarding reliably grading students on a 100-point scale. Still, grading was not always done at schools and grading systems varied widely (Schneider and Hutt, 2013).

By the early 1900s, 100-point or percentage-based grading systems were very common (Cureton, 1971). This period also saw an increased desire for uniformity in grading, and many expressed concerns about what grades meant from one teacher or institution to the next (Weld, 1917). Numerous studies of the period sought to understand and perfect grading systems (Cureton, 1971). Grading on a 100-point scale was found to be highly unreliable, with different teachers unable to assign consistent grades on papers in English, math, and history (Starch, 1913). Researchers felt that getting away from a 100-point scale and grading into only five categories (e.g., letter grades) could increase reliability (Finkelstein, 1913, p. 18). While it is unclear exactly when and why “E” grades disappeared from the letter grade scale, it seems possible that this push to use fewer categories resulted in an “A”–“F” scale with no “E” (“F” being retained, since it so clearly stood for “fail”). Others have conjectured that “E” was removed so students would not assume “E” stood for “excellent,” but whatever the reason, “E’s” apparently disappeared by the 1930s (Palmer, 2010).

As research on intellectual ability appeared to show that, like other continuous biological traits, levels of aptitude in a population conformed to a normal curve, some experts felt grades should similarly be distributed according to a curve in a classroom (Finkelstein, 1913). Distributing grades according to a normal curve was therefore considered as a solution to the subjective nature of grading and a way to minimize inter-rater differences in grading (Guskey, 1994). Others worried that measuring aptitude was different from measuring levels of classroom performance, which might not be normally distributed (Schneider and Hutt, 2013).

Based on the above research and the pressure toward uniformity of grading systems, by the 1940s the “A”–“F” grading system was dominant, with the four-point scale and percentages still also in use (Schneider and Hutt, 2013). However, many inconsistencies remained. As one example, Yale used no less than four different grading systems from the 1960s to 1980s (Yale University, 2013).

Present Day

Grading systems remain controversial and hotly debated today (Jaschik, 2009). Some argue grades are psychologically harmful (Kohn, 1999). Others raise concerns about the integrity of the “A”–“F” system, given well-documented trends in grade inflation (Rojstaczer and Healy, 2012). One professor summed it up by saying grades do no more than “create a facade of coherence” (Jaschik, 2009). A number of colleges have abandoned numerical and categorical grading altogether, opting instead for creating contracts with students to define success or employing student self-reflection in combination with written evaluations by faculty (Jaschik, 2009). Among the Ivy League schools, Brown University does not calculate grade point averages, does not use “D’s” in its grading scale, and does not record failing grades (Brown University, 2014). Even Yale, the institution that started this history of grading more than 200 yr ago, is today still considering changes to its grading system (Yale University, 2013).

Though grades were initially meant to serve various pedagogical purposes, more recent reforms have focused on “grades as useful tools in an organizational rather than pedagogical enterprise—tools that would facilitate movement, communication, and coordination” (Schneider and Hutt, 2013). So, what are the potential purposes of grading in educational settings?

PURPOSES OF GRADING—PAST AND PRESENT

Grades as Feedback on Performance—Does Grading Provide Feedback to Help Students Understand and Improve upon Their Deficiencies?

[This] work affirms an observation that many classroom teachers have made about their students: if a paper is returned with both a grade and a comment, many students will pay attention to the grade and ignore the comment.

—Brookhart (2008, p. 8)

For most faculty members, the concept of feedback has at least two applications to the concept of grading. On one hand, grading itself is a form of feedback that may be
useful to students. In addition, in the process of grading student work, faculty members sometimes provide written comments as feedback that students could use to improve their work. Because college students express a desire for feedback (Higgins et al., 2002), faculty members may feel pressured to grade more (rather than facilitating ungraded activities) and to provide more written feedback while grading. Especially in large classes, this can significantly increase workload on faculty (Nicol and Macfarlane-Dick, 2006; Crisp, 2007). But are grades and written comments effective forms of feedback that assist students in achieving conceptual mastery of the subject?

Feedback is generally divided into two categories: evaluative feedback and descriptive feedback. Evaluative feedback, such as a letter grade or written praise or criticism, judges student work, while descriptive feedback provides information about how a student can become more competent (Brookhart, 2008, p. 26). Butler and Nisan (1986) compared the impacts of evaluative feedback, descriptive feedback, and no feedback on student achievement in problem-solving tasks and in “quantitative” tasks (e.g., those requiring quick, timed work to produce a large number of answers). They found that students receiving descriptive feedback (but not grades) on an initial assignment performed significantly better on follow-up quantitative tasks and problem-solving tasks than did students receiving grades or students receiving no feedback. Students receiving grades performed better on follow-up quantitative tasks than students receiving no feedback, but did not outperform those students on problem-solving assignments. In other words, providing evaluative feedback (in this case, grades) after a task does not appear to enhance students’ future performance in problem solving.

While descriptive, written feedback can enhance student performance on problem-solving tasks; reaping those benefits requires students to read, understand, and use the feedback. Anecdotal accounts, as well as some studies, indicate that many students do not read written feedback, much less use it to improve future work (MacDonald, 1991; Crisp, 2007). In one study, less than half of undergraduate medical students even chose to collect the feedback provided on their essays (Sinclair and Cleland, 2007). Other studies suggest that many students do read feedback and consider it carefully but the feedback is written in a way that students do not find useful in improving future work (Higgins et al., 2002). Some studies have further investigated the relationships between grading and descriptive feedback by providing students with both written feedback and grades on assignments. In these cases, the addition of written comments consistently failed to enhance student performance on follow-up tasks (Marble et al., 1978; Butler 1988; Pulfrey et al., 2011). Brookhart (2008, p. 8) concludes, “the grade ‘trumps’ the comment” and “comments have the best chance of being read as descriptive if they are not accompanied by a grade.” Even when written feedback is read, there is widespread agreement that instructor feedback is very difficult for students to interpret and convert into improved future performance (Weaver, 2006).

Grading does not appear to provide effective feedback that constructively informs students’ future efforts. This is particularly true for tasks involving problem solving or creativity. Even when grading comes in the form of written comments, it is unclear whether students even read such comments, much less understand and act on them.

Grading as a Motivator of Student Effort—Does Grading Motivate Students to Learn?

Our results suggest...that the information routinely given in schools—that is, grades—may encourage an emphasis on quantitative aspects of learning, depress creativity, foster fear of failure, and undermine interest.

—Butler and Nisan (1986)

As described in the history of grading above, our current “A”–“F” grading system was not designed with the primary intent of motivating students. Rather, it stemmed from efforts to streamline communication between institutions and diminish the impacts of unreliable evaluation of students from teacher to teacher (Grant and Green, 2013). That is not to say, however, that grades do not have an impact on student motivation and effort. At some point, every instructor has likely experienced desperate petitions from students seeking more points—a behavior that seems to speak to an underlying motivation stimulated by the grading process.

It would not be surprising to most faculty members that, rather than stimulating an interest in learning, grades primarily enhance students’ motivation to avoid receiving bad grades (Butler and Nisan, 1986; Butler, 1988; Crooks, 1988; Pulfrey et al., 2011). Grades appear to play on students’ fears of punishment or shame, or their desires to outcompete peers, as opposed to stimulating interest and enjoyment in learning tasks (Pulfrey et al., 2011). Grades can dampen existing intrinsic motivation, give rise to extrinsic motivation, enhance fear of failure, reduce interest, decrease enjoyment in class work, increase anxiety, hamper performance on follow-up tasks, stimulate avoidance of challenging tasks, and heighten competitiveness (Harter, 1978; Butler and Nisan, 1986; Butler, 1988; Crooks, 1988; Pulfrey et al., 2011). Even providing encouraging, written notes on graded work does not appear to reduce the negative impacts grading exerts on motivation (Butler, 1988). Rather than seeing low grades as an opportunity to improve themselves, students receiving low scores generally withdraw from class work (Butler, 1988; Guskey, 1994). While students often express a desire to be graded, surveys indicate they would prefer descriptive comments to grades as a form of feedback (Butler and Nisan, 1986).

High-achieving students on initial graded assignments appear somewhat sheltered from some of the negative impacts of grades, as they tend to maintain their interest in completing future assignments (presumably in anticipation of receiving additional good grades; Butler, 1988). Oettinger (2002) and Grant and Green (2013) looked specifically for positive impacts of grades as incentives for students on the threshold between grade categories in a class. They hypothesized that, for example, a student on the borderline between a “C” and a “D” in a class would be more motivated to study for a final exam than a student solidly in the middle of the “C” range. However, these studies found only minimal (Oettinger, 2002) or no (Grant and Green, 2013) evidence that grades motivated students to perform better on final exams under these conditions.

This is not to say that classroom evaluation is by definition harmful or a thing to avoid. Evaluation of students in the service of learning—generally including a mechanism for feedback without grade assignment—can serve to enhance learning and motivation (Butler and Nisan, 1986; Crooks,
Grades as a Tool for Comparing Students—Is Grading on a Curve the Fairest Way to Grade?

You definitely compete for grades in engineering; whereas you earn grades in other disciplines ... I have to get one point higher on the test than the next guy so I can get the higher grade.

—Student quoted in Seymour and Hewitt (1997, p. 118)

The concept of grading on a curve arose from studies in the early 20th century suggesting that levels of aptitude, for example as measured by IQ, were distributed in the population according to a normal curve. Some then argued, if a classroom included a representative sample from the population, grades in the class should similarly be distributed according to a normal curve (Finkelnstein, 1913). Conforming grades to a curve held the promise of addressing some of the problems surrounding grading by making the process more scientific and consistent across classrooms (Meyer, 1908). Immediately, even some proponents of curved grading recognized problems with comparing levels of aptitude in the population with levels of classroom achievement among a population of students. For a variety of reasons, a given classroom might not include a representative sample from the general population. In addition, teachers often grade based on a student’s performance or accomplishment in the classroom—characteristics that differ in many ways from aptitude (Finkelnstein, 1913). However, despite the reservations of some teachers and researchers, curved grading steadily gained acceptance throughout much of the 20th century (Schneider and Hutt, 2013).

Grading on a curve is by definition a type of ”norm-referenced” grading, meaning student work is graded based on comparisons with other students’ work (Brookhart, 2004, p. 72). One issue surrounding norm-referenced grading is that it can dissociate grades from any meaning in terms of content knowledge and of learning. Bloom (1968) pointed out that, in grading on a curve “it matters not that the failures of one year performed at about the same level as the C students of another year. Nor does it matter that the A students of one school do about as well as the F students of another school.” As this example demonstrates, under curved grading, grades might not communicate any information whatsoever regarding a student’s mastery of course knowledge or skills.

Of even more concern, however, is the impact norm-referenced grading has on competition between students. The quote at the start of this section describes how many students respond to curve-graded classes compared with classes that do not use a grading curve. Seymour and Hewitt (1997, p. 118) explain, “Curve-grading forces students to compete with each other, whether they want to or not, because it exaggerates very fine degrees of differences in performance. Where there is little or no difference in work standards, it encourages a struggle to create it.” Studies have shown that science students in competitive class environments do not learn or retain information as well as students in cooperative class environments (Humphreys et al., 1982). Students in cooperative environments are additionally more interested in learning and find learning more worthwhile than students in competitive environments (Humphreys et al., 1982). Of particular concern is that the competitive environment fostered by norm-referenced grading represents one of the factors contributing to the loss of qualified, talented, and often underrepresented college students from science fields (Seymour and Hewitt, 1997; Tobias, 1990). Disturbingly, even when a science instructor does not grade on a curve, students might, due to their past experiences, assume a curve is used and adopt a competitive stance anyway (Tobias, 1990, p. 23).

Bloom (1968, 1976) presents evidence and a theoretical framework supporting an alternate view of grading whereby most students would be expected to excel and not fall into the middle grades. He states, “If the students are normally distributed with respect to aptitude, but the kind and quality of instruction and the amount of time available for learning are made appropriate to the characteristics and needs of each student, the majority of students may be expected to achieve mastery of the subject. And, the relationship between aptitude and achievement should approach zero” (Bloom, 1968). In other words, even if we were to accept a concept of innate aptitude that is normally distributed in a classroom, that distribution should not predict classroom achievement, provided the class environment supports diverse learners in appropriate ways. This idea was a significant development, because it freed teachers from the stigma associated with awarding a larger number of high grades. Previously, an excess of higher grades was thought to arise only from either cheating by students or poor grading practices by teachers (Meyer, 1908). Bloom’s model argues that, when given the proper learning environment and compared against standards of mastery in a field (rather than against one another), large numbers of students could succeed. This type of grading—where instructional goals form the basis of comparison—is called “criterion-referenced” grading (Brookhart, 2004, p. 72).

Of course, Bloom’s work did not rule out the possibility that some teachers might still give high grades for undesirable reasons unrelated to standards of mastery (e.g., to be nice, to gain the admiration of students, etc.). Such practices would not be in line with Bloom’s work and would lead to pernicious grade inflation. Indeed, many of those bemoaning recent trends in grade inflation in higher education (though less prevalent in the sciences) point to the abandonment of curved grading as a major factor (Rojstaczer and Healy, 2012). Such studies often promote various forms of curving—at the level of individual courses or even at the institution as a whole—to combat inflation (Johnson, 2003, chaps. 7–8). In light of the above, however, it seems strange to aspire to introduce grading systems that could further push students into competition and give rise to grades that indicate little about the mastery of knowledge or skills in a subject. The broader
distribution of grades under curve-adjusted grading could simply create the illusion of legitimacy in the grading system without any direct connection between grades and achievement of learning goals. Perhaps the more productive route is to push for stronger, criterion-referenced grading systems in which instructional goals, assessments, and course work are more intimately aligned.

In brief, curved grading creates a competitive classroom environment, alienates certain groups of talented students, and often results in grades unrelated to content mastery. Curving is therefore not the fairest way to assign grades.

**Grades as an Objective Evaluation of Student Knowledge—Do Grades Provide Reliable Information about Student Learning?**

As evidenced by the above headline, some have criticized grading as subjective and inconsistent, meaning that the same student could receive drastically different grades for the same work, depending on who is grading the work and when it is graded. The literature indeed indicates that some forms of assessment lend themselves to greater levels of grading subjectivity than others.

Scoring multiple-choice assessments does not generally require the use of professional judgment from one paper to the next, so instructors should be able to score such assessments objectively (Wainer and Thissen, 1993; Anderson, 2008, p. 451). However, despite their advantages in terms of objective grading, studies have raised concerns regarding the blanket use of multiple-choice assessments. Problems with such assessments range from their potential to falsely indicate student understanding to the possibilities that they hamper critical thinking and exhibit bias against certain groups of students (Townes and Robinson, 1993; Scouller, 1998; Rogers and Harley, 1999; Paxton, 2000; Dufresne et al., 2002; Zimmerman and Williams, 2003; Stanger-Hall, 2012).

Grading student writing, whether in essays, reports, or constructed-response test items, opens up greater opportunities for subjectivity. Shortly after the rise in popularity of percentage-based grading systems in the early 1900s, researchers began examining teacher consistency in marking written work by students. Starch and Elliott (1912) asked 142 teachers to grade the same English paper and found that grades on the paper varied from 50 to 98% between teachers. Because different teachers awarded scores ranging from failing to exceptional, the researchers concluded “the promotion or retardation of a pupil depends to a considerable extent upon the subjective estimate of his teacher” rather than upon the actual work produced by the student (Starch and Elliott, 1912). Even greater levels of inconsistency were found in teachers’ scoring of a geometry paper showing the solution to a problem (Starch and Elliott, 1913).

Eells (1930) investigated the consistency of individual teachers’ grading by asking 61 teachers to grade the same history and geography papers twice—the second time 11 wk after the first. He concluded that “variability of grading is about as great in the same individual as in groups of different individuals” and that, after analysis of reliability coefficients, assignment of scores amounted to “little better than sheer guesses” (Eells, 1930). Similar problems in marking reliability have been observed in higher education environments, although the degree of reliability varies dramatically, likely due to differences in instructor training, assessment type, grading system, and specific topic assessed (Meadows and Billington, 2005, pp. 18–20). Factors that occasionally influence an instructor’s scoring of written work include the penmanship of the author (Bull and Stevens, 1979), sex of the author (Spear, 1984), ethnicity of the author (Fajardo, 1985), level of experience of the instructor (Weigle, 1999), order in which the papers are reviewed (Farrell and Gilbert, 1960; Spear, 1996), and even the attractiveness of the author (Bull and Stevens, 1979).

Designing and using rubrics to grade assignments or tests can reduce inconsistencies and make grading written work more objective. Sharing the rubrics with students can have the added benefit of enhancing learning by allowing for feedback and self-assessment (Jonsson and Svingby, 2007; Reddy and Andrade, 2010). Consistency in grading tests can also be improved by writing longer tests with more narrowly focused questions, but this would tend to limit the types of questions that could appear on an exam (Meadows and Billington, 2005).

In summary, grades often fail to provide reliable information about student learning. Grades awarded can be inconsistent both for a single instructor and among different instructors for reasons that have little to do with a students’ content knowledge or learning advances. Even multiple-choice tests, which can be graded with great consistency, have the potential to provide misleading information on student knowledge.

**GRADING—STRATEGIES FOR CHANGE**

In part, grading practices in higher education have been driven by educational goals such as providing feedback to students, motivating students, comparing students, and measuring learning. However, much of the research literature on grading reviewed above suggests that these goals are often not being achieved with our current grading practices. Additionally, the expectations, time, and stress associated with grading may be distracting instructors from integrating other pedagogical practices that could create a more positive and effective classroom environment for learning. Below we explore several changes in approaching grading that could assist instructors in minimizing its negative influences. Kitchen et al. (2006) additionally provide an example of a high-enrollment college biology class that was redesigned to “maximize feedback and minimize the impact of grades.”

**Balancing Accuracy-Based Grading with Effort-Based Grading**

Multiple research studies described above suggest that the evaluative aspect of grading may distract students from a focus on learning. While evaluation will no doubt always be key in determining course grades, the entirety of students’ grades need not be based primarily on work that rewards only correct answers, such as exams and quizzes. Importantly, constructing a grading system that rewards students for participation and effort has been shown to stimulate...
student interest in improvement (Swinton, 2010). One strategy for focusing students on the importance of effort and practice in learning is to provide students opportunities to earn credit in a course for simply doing the work, completing assigned tasks, and engaging with the material. Assessing effort and participation can happen in a variety of ways (Bean and Peterson, 1998; Rocca, 2010). In college biology courses, clicker questions graded on participation and not correctness of responses is one strategy. Additionally, instructors can encourage students to turn in minute papers in response to a question posed in class and reward this effort based on submission and not scientific accuracy. Perhaps most importantly, biology instructors can assign out-of-class work—case studies, concept maps, and other written assignments—that can promote student practice and focus students’ attention on key ideas, while not creating more grading work for the instructor. Those out-of-class assignments can be graded quickly (and not for accuracy) based on a simple rubric that checks whether students turned the work in on time, wrote the required number of words, posed the required number of questions, and/or included a prescribed number of references. In summary, one strategy for changing grading is to balance accuracy-based grading with the awarding of some proportion of the grade based on student effort and participation. Changing grading in this way has the potential to promote student practice, incentivize in-class participation, and avoid some of the documented negative consequences of grading.

Providing Opportunities for Meaningful Feedback through Self and Peer Evaluation

Instructors often perceive grading to be a separate process from teaching and learning, yet well-crafted opportunities for evaluation can be effective tools for changing students’ ideas about biology. Nicol and Macfarlane-Dick (2006) argue that, just as teaching strategies are shifting away from an instructor-centered, transmissionist approach to a more collaborative approach between instructor and students, so too should classroom feedback and grading. Because feedback traditionally has been given by the instructor and transmitted to students, Nicol and Macfarlane-Dick argue that students have been deprived of opportunities to become self-regulated learners who can detect their own errors in thinking. They advocate for incorporating techniques such as self-reflection and student dialogue into the assessment process. This, they hypothesize, would create feedback that is relevant to and understood by students and would release faculty members from some of the burden of writing descriptive feedback on student submissions. Additionally, peer review and grading practices can be the basis of in-class active-learning exercises, guided by an instructor-developed rubric. For example, students may be assigned out of class homework to construct a diagram of the flow of a carbon atom from a dead body to a coyote (Ebert-May et al., 2003). With the development of a simple rubric, students can self- or peer-evaluate these diagrams during the next class activity to check for the inclusion of key processes, as determined by the instructor. The use of in-class peer evaluation thus allows students to see other examples of biological thinking beyond their own and that of the instructor. In addition, self-evaluation of one’s own work using the instructor’s rubric can build metacognitive skills in assessing one’s own confusions and making self-corrections. Such evaluations need not take much time, and they have the potential to provide feedback that is meaningful and integrated into the learning process. In summary, both self- and peer-evaluation of work are avenues for providing meaningful feedback without formal grading on correctness that can positively influence students’ learning (Sadler and Good, 2006; Freeman et al., 2007; Freeman and Parks, 2010).

Making the Move Away from Curving

As documented in the research literature, the practice of grade curving has had unfortunate and often unintended consequences for the culture of undergraduate science classrooms, pitting students against one another as opposed to creating a collaborative learning community (Tobias, 1990; Seymour and Hewitt, 1997). As such, one simple adjustment to grading would be to abandon grading on a curve. Because the practice of curving is often assumed by students to be practiced in science courses, a move away from curving would likely necessitate explicit and repeated communication with students to convey that they are competing only against themselves and not one another. Moving away from curving practices in grading may remove a key remaining impediment to building a learning community in which students are expected to rely on and support one another in the learning process. In some instances, instructors may feel the need to use a curve when a large proportion of students perform poorly on a quiz or exam. However, an alternative approach would be to identify why students performed poorly and address this more specifically. For example, if the wording of an exam question was confusing for large numbers of students, then curving would not seem to be an appropriate response. Rather, excluding that question from analysis and in computing the exam grade would appear to be a more fair approach than curving. Additionally, if large numbers of students performed poorly on particular exam questions, providing opportunities for students to revisit, revise, and resubmit those answers for some credit would likely achieve the goal of not having large numbers of students fail. This would maintain the criterion-referenced grading system and additionally promote learning of the material that was not originally mastered. In summary, abandoning curving practices in undergraduate biology courses and explicitly conveying this to students could promote greater classroom community and student collaboration, while reducing well-documented negative consequences of this grading practice (Humphreys et al., 1982).

Becoming Skeptical about What Grades Mean

The research literature raises significant questions about what grades really measure. However, it is likely that grades will continue to be the currency of formal teaching and learning in most higher education settings for the near future. As such, perhaps the most important consideration for instructors about grading is to simply be skeptical about what grades mean. Some instructors will refuse to write letters of recommendation for students who have not achieved grades in a particular range in their course. Yet, if grades are not a reliable reflection of learning and reflect other
factors—including language proficiency, cultural background, or skills in test taking—this would seem a deeply biased practice. One practical strategy for making grading more equitable is to grade student work anonymously when possible, just as one would score assays in the laboratory blind to the treatment of the sample. The use of rubrics can also help remove bias from grading (Allen and Tanner, 2006) by increasing grading consistency. Perhaps most importantly, sharing grading rubrics with students can support them in identifying where their thinking has gone wrong and promote learning (Jonsson and Svingby, 2007; Reddy and Andrade, 2010). Much is yet to be understood about what influences students’ performance in the context of formal education, and some have suggested grades may be more of a reflection of a students’ ability to understand and play the game of school than anything to do with learning (Towns and Robinson, 1993; Scoulle, 1998; Stanger-Hall, 2012). In summary, using tools such as rubrics and blind scoring in grading can decrease the variability and bias in grading student work. Additionally, remembering that grades are likely an inaccurate reflection of student learning can decrease assumptions instructors make about students.

IN CONCLUSION—TEACHING MORE BY GRADING LESS (OR DIFFERENTLY)

A review of the history and research on grading practices may appear to present a bleak outlook on the process of grading and its impacts on learning. However, underlying the less encouraging news about grades are numerous opportunities for faculty members to make assessment and evaluation more productive, better aligned with student learning, and less burdensome for faculty and students. Notably, many of the practices advocated in the literature would appear to involve faculty members spending less time grading. The time and energy spent on grading has been often pinpointed as a key barrier to instructors becoming more innovative in their teaching. In some cases, the demands of grading require so much instructor attention, little time remains for reflection on the structure of a course or for aspirations of pedagogical improvement. Additionally, some instructors are hesitant to develop active-learning activities—as either in-class activities or homework assignments—for fear of the onslaught of grading resulting from these new activities. However, just because students generate work does not mean instructors need to grade that work for accuracy. In fact, we have presented evidence that accuracy-based grading may, in fact, demotivate students and impede learning. Additionally, the time-consuming process of instructors marking papers and leaving comments may achieve no gain, if comments are rarely read by students. One wonders how much more student learning might occur if instructors’ time spent grading was used in different ways. What if instructors spent more time planning in-class discussions of homework and simply assigned a small number of earned points to students for completing the work? What if students themselves used rubrics to examine their peers’ efforts and evaluate their own work, instead of instructors spending hours and hours commenting on papers? What if students viewed their peers as resources and collaborators, as opposed to competitors in courses that employ grade curving? Implementing small changes like those described above might allow instructors to promote more student learning by grading less or at least differently than they have before.

REFERENCES


Bagg LH (1871). Four Years at Yale, New Haven, CT: Charles C. Chatfield.


J. Schinske and K. Tanner


Cells are bustling factories with diverse and prolific arrays of molecular machinery. Remarkably, this machinery self-organizes to carry out the complex biochemical activities characteristic of life. When Watson and Crick published the structure of DNA, they noted that DNA base pairing creates a double-stranded form that provides a means of accurately copying the genetic information. Understanding this link between form and function is important for understanding the basis of any biological activity. At the most basic level, a biomolecule’s function is dictated by its structure. The molecule’s shape and chemical properties facilitate interactions with other molecules and determine its role in the cell. Protein function depends on the precise folding of encoded linear stretches of amino acids into three-dimensional shapes. Protein misfolding can lead to various disease states. In this review, we explore engaging online resources that highlight the connection between structure and function in biomolecules. These resources are particularly relevant for instruction at the advanced high school and undergraduate biology levels.

**THE STRUCTURES OF LIFE**

The “Structures of Life” section of the National Institutes of Health (NIH) National Institute of General Medical Sciences website (http://publications.nigms.nih.gov/structlife/index.html) provides an introduction to biomolecules via several resources. A short movie (5:06 min) begins by demonstrating the connection between structure and function in everyday objects and then in biomolecules (Figure 1). The movie illustrates how knowledge of a protein’s structure enables researchers to identify potential drug targets. Finally, the viewer learns how x-ray crystallography and nuclear magnetic resonance spectroscopy are used to determine protein structure. Text-based materials expand on the topics addressed in the movie; these can be accessed in multiple formats: Web pages, a PDF file, an EPUB, or a free paper booklet. These materials also include vignettes of undergraduates, graduate students, and researchers who are studying different aspects of structural biology, capturing some of the diversity of individuals and research in this field. Both the movie and text-based materials focus primarily on protein structure.

**THE PROTEIN DATABASE (PDB)**

The Research Collaboratory for Structural Bioinformatics (RCSB) maintains an international repository of three-dimensional structures for biomolecules called the Protein Database (PDB), available at www.pdb.org. Structural
information in the form of three-dimensional atomic coordinate files (.pdb) is freely available for more than 130,000 biomolecules from hundreds of organisms. To make the PDB more accessible to students and educators, the PDB-101 website provides an entry point to the structural information in the database and supplies complementary educational resources (www.pdb.org/pdb/101/structural_view_of_biology.do). PDB-101 is largely organized around David S. Goodsell’s visually striking “Molecule of the Month” feature, which highlights the structure and function of a molecule, usually a protein, with high-quality illustrations based on PDB structural data. For example, Goodsell’s illustrations of ferritin effectively depict how protein subunits fit together to form a porous shell to uptake and store iron (Figure 2). “Molecule of the Month” addresses the importance of the featured molecule in cellular, tissue, organismal, and health contexts. Although the content is very detailed and a basic understanding of molecular biology is assumed, the language is approachable and contains minimal jargon. Because structural information about a molecule often comes from a variety of organisms and in a variety of conditions or conformations, Goodsell points users toward pertinent structural data files. Embedded links to these interactive PDB files and a short list of relevant references encourage readers to explore the structural data and literature. The “Molecule of the Month” content can also be accessed from PDB-101’s “Structural View of Biology” interface, which nicely organizes the featured molecules (currently more than 80) by functional category.

PDB-101’s “Educational Resources” tab contains structural posters, paper model tutorials, lesson ideas, and animations. Among the animations, What Is a Protein? (3:39 min) is a useful introduction to protein structure and the relationship between form and function. The movie introduces amino acid chemistry and the four levels of protein structure. An animation illustrates how the structure of hemoglobin enables heme binding and oxygen transport. The connection between protein form and function is also highlighted for proteins involved in immunity, nutrient storage, cytoskeleton, enzymatic activity, cellular communication, and transport. The molecular depictions are based on published structural data, and the animations effectively depict molecular interactions and conformational changes. While not integral to the accuracy of the information presented, a less repetitive soundtrack and narration by a professional voice-over artist would enhance the quality of the movie.

Finally, PDB-101’s “Understanding PDB Data” section outlines how PDB structural data are generated and how these data files can be downloaded, visualized, and manipulated. This section also outlines the three major types of molecular diagrams: 1) wire frame, in which covalent bonds are depicted as lines; 2) space filling, in which each atom is represented as an electron cloud; and 3) backbone and ribbon, which illustrates secondary and tertiary structures. Each type of diagram is useful for visualizing different aspects of a molecule’s structure.

**INTERACTING WITH MOLECULAR MODELS**

Several freely available programs are available for working with and visualizing atomic coordinate files from the PDB. One of these, PyMOL, is freely available for educational use at http://pymol.org. Jmol: An Open-source Java Viewer for Chemical Structures in 3D (www.jmol.org) is an applet designed for use in Web pages. Websites that run Jmol require that a user’s computer have the latest version of the Java software (www.java.com/en) and that Java applets are enabled in its Web browser; no other software is required. FirstGlance in Jmol (http://bioinformatics.org/firstglance/fjgj) is a simple platform with sophisticated functionality for viewing a molecule’s structure with different diagrams, cross-sections, and emphasis on various molecular features. Currently, many sites, including the PDB, are converting from Jmol applets to

---

**Figure 2.** PDB-101’s “Molecule of the Month” illustration of the iron storage protein ferritin reveals how assembled protein subunits form a hollow sphere with pores for iron internalization. Illustration by David Goodsell, the Scripps Research Institute and RCSB PDB. (www.rcsb.org/pdb/101/motm.do?momID=35)
the HTML5/Javascript-based JSmol, which works on tablets (e.g., iPads). The Web-based Jmol and JSmol also allow for the construction and exploration of online tutorials that guide the user through the visualization of a molecule.

An excellent source for molecular visualization tutorials is the Online Macromolecular Museum (OMM; www.callutheran.edu/Academic_Programs/Departments/BioDev/omm/exhibits.htm) maintained by David Marcey at California Lutheran College. Marcey has undergraduate students delve into the scientific literature and structural data to author molecular tutorials, which he curates and adds to OMM. The website also explains how users can build their own tutorials and includes information on working with Jmol script. OMM is organized by topic area into various “halls” (e.g., DNA/RNA, virology, immunology, signal transduction), each containing “exhibits” that step users through exploration of a molecular structure. Some tutorials use the outdated molecular visualization tool Chime, which requires its own plug-in (www.umass.edu/microbio/chime/getchime.html#download), but these are being phased out in favor of Jmol and JSmol. OMM tutorials provide a view of the molecule on the left and the tutorial text with functional radio buttons on the right (Figure 3). Each exhibit provides useful Web browser and tutorial navigation tips. An introduction to the molecule is provided at the beginning of the tutorial, but some knowledge of biology and types of molecular diagrams is a prerequisite. The tutorial guides users with a series of radio buttons that load new structures, animate molecules, or alter the molecule’s display to highlight specific features. Animations of conformational changes and binding interfaces between molecules connect details of a molecule’s structure to its function. The user also can rotate the molecule by clicking and dragging to engage with the three-dimensional form. The inclusion of load/reset molecule buttons is useful here. Exhibits conclude with a list of references for further study. Because of the high level of detail, the exhibits work best at the undergraduate level, as supplements to a textbook or classroom instruction.

They also lend themselves to self-guidance, so users can proceed at their own pace and replay animations. Large volumes of structural data can make the tutorials slow to load.

Other sites that include tutorials exploring molecular structures include the collaborative wiki Proteopedia (www.proteopedia.org/wiki/index.php/Main_Page), Milwaukee School of Engineering’s Center for BioMolecular Modeling (http://cbm.msoe.edu), and Eric Martz’s MolviZ (www.molviz.org).

**ANIMATING STRUCTURE AND FUNCTION**

The expanding collection of structural information has given rise to animations that capture how structural features lead to function in a cellular context. The Howard Hughes Medical Institute (HHMI) and the Dolan DNA Learning Center (DNALC) collaborated on the creation of animations

**Figure 3.** The Online Macromolecular Museum’s exhibit on actin shows how the two major domains of an actin protein form a cleft in which ATP binds. David Marcey, California Lutheran College. (www.callutheran.edu/BioDev/omm/jmolxx/myosin_actin/myosin_actin.html)

**Figure 4.** The HHMI and Dolan DNALC’s Mechanism of Replication animation uses structural data to animate how the protein machinery unwinds, separates, and copies each strand of DNA. Copyright HHMI. (www.hhmi.org/biointeractive/dna-replication-advanced-detail and www.dnalc.org/resources/3d/04-mechanism-of-replication-advanced.html)
(ranging in length from 39 s to 3:04 min) that harness knowledge of structural information to show molecular function. These molecular animations are available through HHMI’s BioInteractive site (www.biointeractive.org/animations) and the Dolan DNALC’s three-dimensional animation library (www.dnalc.org/resources/3d). Although there are multiple ways to search for the animations on the HHMI BioInteractive site, the most direct approach is to access all of their animations via the above URL and then select “DNAi” under “Series” in the left-hand column on the Web page. The URL for the Dolan DNALC website provides a direct link to the animations. Each site also contains unique content that is worth exploring.

Four of the animations (How DNA Is Packaged, Transcription, Translation, and DNA Replication) are available with both a basic and an advanced level of narration; the basic level of How DNA Is Packaged is only available on the Dolan DNALC site. For example, the Mechanism of Replication animations (Figure 4) show how DNA’s double-stranded structure serves as a dual template. Both the basic and advanced versions capitalize on resolved structural information to show the proteins that facilitate DNA replication and how one strand is copied “backward” due to the antiparallel nature of the DNA strands. In the basic version (1:07 min), viewers see how molecules fit together to form the complex replication machinery, which underscores the importance of molecular structure. The advanced version (2:20 min) goes into greater detail about individual components of the replication process, adds labels, and focuses on specific features of DNA replication.

As the resources reviewed here illustrate, the field of structural biology is generating a wealth of information linking molecular form and function. Helping students build these connections can make complicated molecular phenomena more accessible.

ACKNOWLEDGMENTS
We thank A. Malcolm Campbell for critical comments and helpful suggestions on this feature.
INTRODUCTION

When we were first invited to write these columns, the editors felt it would be an interesting way to give the readers of *CBE—Life Sciences Education* an agency’s-eye view of its concerns, workings, and accomplishments. This column is written with that charge in mind. It is intended to inform the community about our outreach efforts, some of the mechanisms used, and the effectiveness of these efforts.

In 2007, the National Science Foundation (NSF), the American Association for the Advancement of Science (AAAS), the Howard Hughes Medical Institute (HHMI), and the National Institutes of Health (NIH) launched an initiative called Vision and Change in Undergraduate Biology Education (V&C), designed to enable the biology community to address the challenges of preparing undergraduates for 21st-century biology using 21st-century findings about effective teaching strategies. Over the next 2 yr, with support from the partners listed above, the community engaged in “conversations” among faculty, administrators, students, and professional societies through a series of workshops (http://visionandchange.org/files/2010/03/VC-Preliminary-Reports-from-Conversations1.pdf).

Those conversations culminated in a working conference in 2009, in which more than 500 faculty, administrators, and students pooled their collective wisdom (visionandchange.org/about). The tangible result from all this effort, *Vision and Change in Undergraduate Biology Education: A Call to Action* (http://visionandchange.org/files/2011/03/Revised-Vision-and-Change-Final-Report.pdf), is a document setting forth their vision of the approaches needed to ensure undergraduate biology education truly reflects the biology of the 21st century in both the content included and the competencies developed and the delivery reflects what is known about effective teaching practices. Its message is summarized in four action statements “aimed at ensuring that the vision of the conference becomes an agenda for change” (Vision and Change Final Report; http://visionandchange.org/files/2011/03/Revised-Vision-and-Change-Final-Report.pdf):

- Integrate core concepts and competencies throughout the curriculum
- Focus on student-centered learning
- Promote a campus-wide commitment to change
- Engage the biology community in the implementation of change

Since the release of the report, many organizations, including those listed above, the U.S. Department of Agriculture, the National Research Council, and a number of professional societies have led efforts to disseminate the document and engage the biology community in efforts to implement the actions outlined in it. These efforts have recently been enhanced by such initiatives as the Partnership for Undergraduate Life Sciences Education (PULSE; www.pulsecommunity.org) and CourseSource (http://coursesourcejournal.wordpress.com), both designed with the specific aim of disseminating the ideas presented in V&C and catalyzing their widespread implementation.1,2

In addition, a second V&C conference, *Vision and Change:...*
ATransforming the Change, was held in August 2013 to document and encourage adoption of the key ideas at the departmental and institutional levels.

How successful have these efforts been? Six years after those initial conversations, it is important to try to determine the extent of dissemination and implementation of V&C, not only to determine the challenges ahead and possible next steps, but also to begin to establish mechanisms for effective response when a science experiences the sort of rapid evolution in approach, resources, and focus that has occurred recently in biology.

Finding interesting and informative anecdotes concerning individual efforts is relatively easy; documenting adoption on a more universal scale within the discipline is a much more formidable challenge. This paper summarizes the results of efforts to systematically analyze the impact of V&C by examining the references and ideas presented in proposals submitted to various programs within the NSF, documenting use of V&C in the community through Web data analytics, and investigating the presentation and discussion of V&C at professional society events. It also includes initial findings reported at the August 2013 conference. We hope both to inform the reader and to inspire reflection within the readership about effective means of determining the outcomes of efforts such as V&C that aim to meet the broad general needs of a science, technology, engineering, and mathematics (STEM) community.

For example, preliminary evidence, gleaned from analysis of biology education–related proposals submitted to a number of NSF programs, indicates that although there is good evidence of use by some segments of the biology community, there is less use of V&C principles or even acknowledgment of its existence within other segments (see tables and figures). This suggests the need to develop dissemination efforts that resonate well with these communities.

### EVIDENCE WITHIN NSF

The Transforming Undergraduate Education in STEM (TUES) Program, a program offered from 2010 to 2013, sought to improve the quality of STEM education for all undergraduate students (www.nsf.gov/pubs/2010/nsf10544/nsf10544.pdf). TUES included both pilot projects (TUES type 1) and those with a more global reach at the departmental, institutional, or national level (TUES types 2 and 3). This program was a natural home for those seeking support for V&C-related projects. Therefore, biology-related TUES proposals from 2009 (the year of the first V&C meeting) to 2013 were examined with two goals in mind: 1) to ascertain the growth of community awareness of the V&C initiative and 2) to determine the number of proposals whose project design and implementation reflected the recommendations of V&C.

To determine the growth of community awareness of V&C, we examined proposals and documented the number that cited V&C at least once within the proposal (Table 1). The data in Table 1 show that < 1% of the proposals submitted to the TUES type 1 track in 2009 included a reference to V&C. In each subsequent year, the percentage of proposals including a reference to V&C increased dramatically, indicating increasing knowledge of the existence of V&C by TUES submitters (faculty in a variety of fields of biology and from a variety of institutions). The trend is similar for all types of TUES proposals, ranging from pilot project submissions (type 1) to proposals with a wider range of influence (types 2 and 3).

The increasing frequency of citations in TUES proposals over time suggests that V&C is recognized as a significant report to be cited as part of the required literature base for proposals to the program, but it does not reveal the extent to which the recommendations of V&C were drivers of project design. To study the influence of V&C on the design of the TUES projects, we examined the language used in the proposals. It was determined that the terms core competencies, student-centered, and core concept correlated with the use of the recommendations of V&C in the design of the project. Any proposal that used two or more of these terms was scored as having a V&C-influenced design (Figure 1).

The data in Figure 1 suggest that the V&C report is not only an important element of the literature base for proposals seeking to improve undergraduate biology education, it is increasingly the foundation for the design of those projects. While this is a preliminary analysis, it is encouraging that V&C is having an increasing impact among the community of biology educators who seek external funding from NSF to improve undergraduate biology education at their home institutions.

We next examined the impact of V&C on other NSF programs that receive biology-specific proposals for education projects. The data show that not all undergraduate education programs see the same integration of V&C principles and suggest that action is needed to disseminate the recommendations more effectively. For instance, in the Advanced Technology Education Program biology proposals, the percentage of references to V&C increased from 0 to ∼9% between

---

**Table 1. Analysis of the frequency of V&C citations in TUES biology-related proposals from 2009–2013**

<table>
<thead>
<tr>
<th>TUES date of proposal submission</th>
<th>Number of proposals submitted</th>
<th>Number of proposals citing V&amp;C</th>
<th>Percent of proposals citing V&amp;C</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009, May type 1</td>
<td>108</td>
<td>1</td>
<td>0.93</td>
</tr>
<tr>
<td>2010, May type 1</td>
<td>98</td>
<td>14</td>
<td>15.3</td>
</tr>
<tr>
<td>2011, May type 2</td>
<td>79</td>
<td>27</td>
<td>34.2</td>
</tr>
<tr>
<td>2012, May type 1</td>
<td>98</td>
<td>44</td>
<td>44.9</td>
</tr>
<tr>
<td>2010, January types 2 and 3</td>
<td>17</td>
<td>2</td>
<td>11.76</td>
</tr>
<tr>
<td>2011, January types 2 and 3</td>
<td>28</td>
<td>16</td>
<td>57.1</td>
</tr>
<tr>
<td>2012, January types 2 and 3</td>
<td>18</td>
<td>12</td>
<td>66.7</td>
</tr>
<tr>
<td>2013, January types 2 and 3</td>
<td>32</td>
<td>19</td>
<td>59.4</td>
</tr>
</tbody>
</table>

*aData from the NSF report server.*
Figure 1. Biology-specific proposals submitted to the Course, Curriculum, and Laboratory Improvement/TUES Program in the years 2009–2013, using only those submitted post-February 2009 (the date of the V&C meeting), were analyzed for V&C citations and for evidence that the design of the program reflected principles within V&C. The proposal numbers are aggregates of type 1 and type 2/3 submissions and are grouped according to the calendar year in which they were submitted. The percentages listed above the bars reflect percentage of the total number of proposals analyzed in that year. The low proposal number in 2013 reflects the fact there was no TUES 1 competition in that year. Data from the NSF report server.

2009 and 2012 (Table 2). This program is primarily focused on training technicians in community colleges. It suggests that the V&C initiative has not been as successful in engaging faculty in the 2-yr schools as it has been in other institutions of higher education. Among the biology proposals submitted to the Tribal Colleges and Universities Program (T-CUP) and the Historically Black Colleges and Universities Undergraduate Program (HBCU-UP), there is only one proposal that references V&C. T-CUP and HBCU-UP programs are designed to improve STEM education and preparation of undergraduate

Table 2. Analysis of the frequency of V&C citations from 2008–2012 in proposals submitted to several NSF programsa

<table>
<thead>
<tr>
<th>Date of proposal submission</th>
<th>Number of proposals sampled</th>
<th>Number of proposals citing V&amp;C</th>
<th>Percent of proposals citing V&amp;C</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATE, biology-identified proposalsb</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>23</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>2010</td>
<td>36</td>
<td>1</td>
<td>2.8</td>
</tr>
<tr>
<td>2011</td>
<td>22</td>
<td>2</td>
<td>9.1</td>
</tr>
<tr>
<td>T-CUP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>7</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>2010</td>
<td>15</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>2011</td>
<td>15</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>HBCU-UP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>3</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>2010</td>
<td>26</td>
<td>1</td>
<td>3.8</td>
</tr>
<tr>
<td>2011</td>
<td>45</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>CAREER biology awards</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>83</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>2009</td>
<td>55</td>
<td>1</td>
<td>1.8</td>
</tr>
<tr>
<td>2010</td>
<td>46</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>2011</td>
<td>66</td>
<td>4</td>
<td>6.1</td>
</tr>
<tr>
<td>2012</td>
<td>44</td>
<td>3</td>
<td>6.8</td>
</tr>
</tbody>
</table>

aData from the NSF report server.
bBiology-focused proposals that have been submitted to the ATE program.
students in these minority-serving institutions, suggesting that V&C outreach to faculty in these institutions needs to be more vigorous.

The increase in V&C citations in the biology proposals to the Faculty Early Career Development (CAREER) Program has been modest, rising from ∼2% in 2009 to nearly 7% in 2012 (Table 2). Because this program supports junior faculty who “exemplify the role of teacher-scholars through outstanding research, excellent education and the integration of education and research within the context of the mission of their organizations,” greater awareness of V&C would have been expected (CAREER Solicitation; www.nsf.gov/pubs/2014/nsf14532/nsf14532.htm).

EVIDENCE IN THE COMMUNITY

Looking beyond NSF, changes in biology community awareness of V&C were examined by analyzing Web data referencing V&C. From October 2011 to September 2012, each instance of “Vision and Change in Undergraduate Biology Education” on the Web was catalogued through the use of Web search engines. Data regarding the host of the material, the date posted, and the type of data were recorded to develop a broader understanding of V&C in the academic community as manifested online.

While generating a snapshot over time, the heterogeneity of the data, while intriguing, made it difficult to assess in a comprehensible way. Consequently, we chose to examine the instances of V&C that occurred in journal articles, in postings by universities and colleges, and in postings by professional societies, as for faculty these have clear associations and meaning for the discipline. The data were restricted to instances in which clear identification of the date posted and the identity of the host were possible, greatly reducing the sample sizes (Figure 2).

The data show increasing numbers of Internet citations in the period 2009–2012 (Figure 2). This trend mirrors the observations from the proposal investigations and further suggests there is increasing awareness and penetration of the V&C report within the academic community. Furthermore, as professional societies have integral roles in the discipline, detailed agendas of recent and future annual meetings were analyzed. Of the 40 societies examined, 23 provided information regarding V&C, while 11 had V&C-related talks at their annual meeting.

The AAAS was excluded from the collection of professional society data, because it hosts the official V&C website (www.visionandchange.org). However, the increased usage of that site is also an indicator of the community’s responsiveness to the V&C initiative. Since its inception in July 2010, the site has had 36,302 visitors, with average visitors per month increasing each year from 334 visits per month in 2010 to 1338 visits per month in 2013. Of these, 65% were new visitors and 35% were repeat visitors. It is interesting to note that 80% of the visits were from individuals typing “vision and change” into the Web browser or searching with a search engine. This suggests that these visitors were familiar with the report and were seeking additional information. Ten percent of the visits were referred by other websites, including AAAS, the American Society for Microbiology, the American Institute of Biological Sciences, CBE—Life Sciences Education, and the American Society of Plant Biologists.

CHRONICLING THE CHANGE

In August 2013, AAAS hosted Vision and Change in Undergraduate Biology: Chronicling the Change, Inspiring the Future (http://visionandchange.org/about-v-c-chronicling-the-changes), a working meeting that drew more than 350 faculty members and administrators, as well as professional society representatives. Like the first meeting in 2009, this meeting had specific goals, including sharing innovative strategies, identifying common challenges, understanding the role of leadership in driving change in undergraduate biology education, and strengthening collaborative networks. The posters submitted included a range of actions from...
individual classroom efforts to department-wide initiatives. Evidence of university and community college system-wide collaborative change and of the development of communities of practice through professional societies suggests that V&C is having a significant impact on undergraduate biology education. A report synthesizing what has been learned, strategies found to be successful, and needed next steps will be released in Spring of 2014.

Additional evidence of the penetration of V&C is the establishment of a variety of resources by the supporting agencies. Organizations such as PULSE, CourseSource, and many biology professional societies are providing a wealth of resources, communities of practice, and professional development opportunities to support faculty in their efforts to implement the curricular changes recommended in V&C. Notable among these initiatives is the PULSE community website, which currently serves nearly 1200 members. The community continues to expand the variety of resources available to faculty, departments, and institutions striving to implement the V&C recommendations.

Supporting V&C has been an interesting effort for the agencies involved as we try to help the community respond to what they and we see as both a unique opportunity and an interesting challenge—the need to respond to the changes in a science that is constantly finding new directions and resources—while at the same time learning to utilize the resources provided by current knowledge about how people learn. It is rather amazing that we have worked together on these issues for almost 8 yr, and you, the community of biology educators, have kept us going. We have been informed, inspired, and energized by the community response to our calls for action.

Our efforts to catalogue changes made are just beginning, but seem to have made progress. We depend on you for suggestions for mechanisms we can use to get a sense of community actions. Information about individual faculty, institutions, or professional societies is not always easy to catalogue, but this information is simple compared with getting a sense of the community as a whole. We hope this article triggers ideas from the community as to how to capture the larger picture. Armed with this accumulated knowledge, each segment of the community can move forward with a clearer understanding of the next steps necessary and the mechanisms to use to transform undergraduate biology education and prepare students for the 21st century. We hope that this article will spark ideas we can use to further chronicle the changes. Send your reactions and ideas to us at kdennist@nsf.gov or hvasaly@nsf.gov. Thank you!

ACKNOWLEDGMENTS

We thank Terry Woodin and Cynthia Bauerle for assistance and for comments on the manuscript. We thank Sarah Elgin for her insightful and quick editing.
Assessment of Course-Based Undergraduate Research Experiences: A Meeting Report


*University of Georgia, Athens, GA 30602; †University of Colorado–Boulder, Boulder, CO 80309; ‡University of Wisconsin–Madison, Madison, WI 53706; §University of California, Los Angeles, Los Angeles, CA 90995; ¶Indiana University of Pennsylvania, Indiana, PA 15705; #University of Queensland, Brisbane 4072, Australia; @Cornell University, Ithaca, NY 14850; **Purdue University, West Lafayette, IN 47907; ††Indiana University–Purdue University, Indianapolis, IN 46202

The Course-Based Undergraduate Research Experiences Network (CUREnet) was initiated in 2012 with funding from the National Science Foundation program for Research Coordination Networks in Undergraduate Biology Education. CUREnet aims to address topics, problems, and opportunities inherent to integrating research experiences into undergraduate courses. During CUREnet meetings and discussions, it became apparent that there is need for a clear definition of what constitutes a CURE and systematic exploration of what makes CUREs meaningful in terms of student learning. Thus, we assembled a small working group of people with expertise in CURE instruction and assessment to: 1) draft an operational definition of a CURE, with the aim of defining what makes a laboratory course or project a “research experience”; 2) summarize research on CUREs, as well as findings from studies of undergraduate research internships that would be useful for thinking about how students are influenced by participating in CURES; and 3) identify areas of greatest need with respect to CURE assessment, and directions for future research on and evaluation of CUREs. This report summarizes the outcomes and recommendations of this meeting.

Students can work with the same data at the same time and with the same tools as research scientists.

iPlant Education, Outreach & Training Group (2008, personal communication)

INTRODUCTION

Numerous calls for reform in undergraduate biology education have emphasized the value of undergraduate research experiences (e.g., American Association for the Advancement of Science [AAAS], 2011). These calls are based on a growing body of research that documents how students benefit from research experiences (Kremer and Bringle, 1990; Kardash, 2000; Rauchhorst et al., 2001; Hathaway et al., 2002; Bauer and Bennett, 2003; Lopatto, 2004, 2007; Lopatto and Tobias, 2010; Seymour et al., 2004; Hunter et al., 2007; Russell et al., 2007; Laursen et al., 2010; Thiry and Laursen, 2011). Undergraduates who participate in research internships (also called research apprenticeships, undergraduate research experiences, or research experiences for undergraduates [REUs]) report positive outcomes, such as learning to think like a scientist, finding research exciting, and intending to pursue graduate education or careers in science (Kardash, 2000; Laursen et al., 2010; Lopatto and Tobias, 2010). Research experiences are thought to be especially beneficial for women and underrepresented minority students, presumably because they support the development of relationships with more senior scientists and with peers who can offer critical support to students who might...
of a sense of ownership can contribute to their persistence in science.

The Course-Based Undergraduate Research Experiences Network (CUREnet; http://curenet.franklin.uga.edu) was initiated in 2012 with funding from NSF to support CURE instruction by addressing topics, problems, and opportunities inherent to integrating research experiences into undergraduate courses. During early discussions, the CUREnet community identified a need for a clearer definition of what constitutes a CURE and a need for systematic exploration of how students are affected by participating in CURES. Thus, a small working group with expertise in CURE design and assessment was assembled in September 2013 to:

1. Draft an operational definition of a CURE;
2. Summarize research on CURES, as well as findings from studies of undergraduate research internships that would be useful for thinking about how students are influenced by participating in CURES; and
3. Identify areas of greatest need with respect to evaluation of CURES and assessment of CURE outcomes.

In this paper, we summarize the meeting discussion and offer recommendations for next steps in the assessment of CURES.

CURES DEFINED

The first aim of the meeting was to define a CURE. We sought to answer the question: How can a CURE be distinguished from other laboratory learning experiences? This allows us to make explicit to students how a CURE may differ from their other science course work and to distinguish a CURE from other types of learning experiences for the purposes of education research and evaluation. We began by discussing what we mean by “research.” We propose that CURES involve students in the following:

1. Use of scientific practices. Numerous policy documents, as well as an abundance of research on the nature and practice of science, indicate that science research involves the following activities: asking questions, building and evaluating models, proposing hypotheses, designing studies, selecting methods, using the tools of science, gathering and analyzing data, identifying meaningful variation, navigating the messiness of real-world data, developing and critiquing interpretations and arguments, and communicating findings (National Research Council [NRC], 1996; Singer et al., 2006; Duschl et al., 2007; Bruck et al., 2008; AAAS, 2011; Quinn et al.,

| Table 1. Features of CUREs compared with research internships |
|---------------------------------|-----------------|-----------------|
|                  | CURES            | Research internships |
| Scale             | Many students    | Few students     |
| Mentorship structure | One instructor to many students | One instructor to one student |
| Enrollment         | Open to all students in a course | Open to a selected or self-selecting few |
| Time commitment    | Students invest time primarily in class | Students invest time primarily outside class |
| Setting            | Teaching lab     | Faculty research lab |
Discovery. Discovery is the process by which new knowledge or insights are obtained. Science research aims to generate new understanding of the natural world. As such, discovery in the context of a CURE implies that the outcome of an investigation is unknown to both the students and the instructor. When the outcomes of their work are not predetermined, students must make decisions such as how to interpret their data, when to track down an anomaly and when to ignore it as “noise,” or when results are sufficiently convincing to draw conclusions (Duschl et al., 2007; Quinn et al., 2011). Discovery carries with it the risk of unanticipated outcomes and ambiguous results because the work has not been done before. Discovery also necessitates exploration and evidence-based reasoning. Students and instructors must have some familiarity with the current body of knowledge in order to contribute to it and must determine whether the new evidence gathered is sufficient to support the assertion that new knowledge has been generated (Quinn et al., 2011). We propose that discovery in the context of a CURE means that students are addressing novel scientific questions aimed at generating and testing new hypotheses. In addition, when their work is considered collectively, students’ findings offer some new insight into how the natural world works.

3. Broadly relevant or important work. Because CUREs provide opportunities for students to build on and contribute to current science knowledge, they also present opportunities for impact and action beyond the classroom. In some CUREs, this may manifest as authorship or acknowledgment in a science research publication (e.g., Leung et al., 2010; Pope et al., 2011). In other CUREs, students may develop reports of interest to the local community, such as a report on local water quality or evidence-based recommendations for community action (e.g., Savan and Sider, 2003). We propose that CUREs involve students in work that fits into a broader scientific endeavor that has meaning beyond the particular course context. (We choose the language of “broader relevance or importance” rather than the term “authenticity” because views on the authenticity of a learning experience may shift over time [Rahm et al., 2003] and may differ among students, instructors, and the broader scientific community.)

4. Collaboration. Science research increasingly involves teams of scientists who contribute diverse skills to tackling large and complex problems (Quinn et al., 2011). We propose that group work is not only a common practical necessity but also an important pedagogical element of CUREs because it exposes students to the benefits of bringing together many minds and hands to tackle a problem (Singer et al., 2006). Through collaboration, students can improve their work in response to peer feedback. Collaboration also develops important intellectual and communication skills as students verbalize their thinking and practice communicating biological ideas and interpretations either to fellow students in the same discipline or to students in other disciplines. This may also encourage students’ metacognition—solidifying their thinking and helping them to recognize shortcomings in their knowledge and reasoning (Chi et al., 1994; Lyman, 1996; Smith et al., 2009; Tanner, 2009).

5. Iteration. Science research is inherently iterative because new knowledge builds on existing knowledge. Hypotheses are tested and theories are developed through the accumulation of evidence over time by repeating studies and by addressing research questions using multiple approaches with diverse methods. CUREs generally involve students in iterative work, which can occur at multiple levels. Students may design, conduct, and interpret an investigation and, based on their results, repeat or revise aspects of their work to address problems or inconsistencies, rule out alternative explanations, or gather additional data to support assertions (NRC, 1996; Quinn et al., 2011). Students may also build on and revise aspects of other students’ investigations, whether within a single course to accumulate a sufficiently large data set for analysis or across successive offerings of the course to measure and manage variation, further test preliminary hypotheses, or increase confidence in previous findings. Students learn by trying, failing, and trying again, and by critiquing one another’s work, especially the extent to which claims can be supported by evidence (NRC, 1996; Duschl et al., 2007; Quinn et al., 2011).

These activities, when considered in isolation, are not unique to CUREs. Rather, we propose that it is the integration of all five dimensions that makes a learning experience a CURE. Of course, CUREs will vary in the frequency and intensity of each type of activity. We present the dimensions in Table 2 and delineate how they are useful for distinguishing between the following four laboratory learning environments:

1. A traditional laboratory course, in which the topic and methods are instructor defined; there are clear “cookbook” directions and a predetermined outcome that is known to students and to the instructor (Domin, 1999; Weaver et al., 2008);
2. An inquiry laboratory course, in which students participate in many of the cognitive and behavioral practices that are commonly performed by scientists; typically, the outcome is unknown to students, and they may be challenged to generate their own methods. The motivation for the inquiry is to challenge the students, rather than contribute to a larger body of knowledge (Domin 1999; Olson and Loucks-Horsley, 2000; Weaver et al., 2008);
3. A CURE, in which students address a research question or problem that is of interest to the broader community with an outcome that is unknown both to the students and to the instructor (Domin 1999; Bruck et al., 2008; Weaver et al., 2008); and
### Table 2. Dimensions of different laboratory learning contexts

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Traditional</th>
<th>Inquiry</th>
<th>CURE</th>
<th>Internship</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use of science practices</td>
<td>Few scientific practices</td>
<td>Multiple scientific practices</td>
<td>Multiple scientific practices</td>
<td>Multiple scientific practices</td>
</tr>
<tr>
<td></td>
<td>Instructor driven</td>
<td>Student driven</td>
<td>Student or instructor driven</td>
<td>Student or instructor driven</td>
</tr>
<tr>
<td>Study design and methods</td>
<td>Instructor defined</td>
<td>Student defined</td>
<td>Student or instructor defined</td>
<td>Student or instructor defined</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discovery</td>
<td>Instructor defined</td>
<td>Student defined</td>
<td>Student or instructor defined</td>
<td>Student or instructor defined</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outcome is</td>
<td>Known to students and instructors</td>
<td>Varied</td>
<td>Unknown</td>
<td>Unknown</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Findings are</td>
<td>Previously established</td>
<td>May be novel</td>
<td>Novel</td>
<td>Novel</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Broader relevance or</td>
<td>Students’ work presents opportunities for action</td>
<td>Is limited to the course</td>
<td>Extends beyond the course</td>
<td>Extends beyond the course</td>
</tr>
<tr>
<td>importance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collaboration</td>
<td>Collaboration occurs</td>
<td>Among students in a course</td>
<td>Among students, teaching assistants, instructor in a course</td>
<td>Between student and mentor in a research group</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instructor’s role is</td>
<td>Instruction</td>
<td>Facilitation</td>
<td>Guidance and mentorship</td>
<td>Guidance and mentorship</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iteration</td>
<td>Risk of generating &quot;messy&quot; data are…</td>
<td>Minimized</td>
<td>Significant</td>
<td>Inherent</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Iteration is built into the process</td>
<td>Occasionally</td>
<td>Often</td>
<td>Often</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4. A research internship, in which a student is apprenticed to a senior researcher (faculty, postdoc, grad student, etc.) to help advance a science research project (Seymour et al., 2004).

The five dimensions comprise a framework that can be tested empirically by characterizing how a particular dimension is manifested in a program, developing scales to measure the degree or intensity of each dimension, and determining whether the dimensions in part or as a whole are useful for distinguishing CUREs from other laboratory learning experiences. Once tested, we believe that this framework will be useful to instructors, institutional stakeholders, education researchers, and evaluators.

Instructors may use the framework to delineate their instructional approach, clarify what students will be expected to do, and articulate their learning objectives. For example, in traditional laboratory instruction, students may collect and analyze data but generally do not build or evaluate models or communicate their findings to anyone except the instructor. During inquiry laboratory instruction, students may be able to complete a full inquiry cycle and thus engage at some level in the full range of scientific practices. Students in CUREs and research internships may engage in some scientific practices in depth, but neglect others, depending on the particular demands of the research and the structure of the project. As instructors define how their course activities connect to desired student outcomes, they can also identify directions for formative and summative assessment.

Education researchers and evaluators may use the framework to characterize particular instructional interventions with the aim of determining which dimensions, to what degree and intensity, correlate with desired student outcomes. For instance, students who engage in the full range of scientific practices could reasonably be expected to improve their skills across the range of practices, while students who participate in only a subset of practices can only be expected to improve in those specific practices. Similarly, the extent to which students have control over the methods they employ may influence their sense of ownership over the investigation, thus increasing their motivation and perhaps contributing to their self-identification as scientists. Using this framework to identify critical elements of CUREs and how they relate (or not) to important student outcomes can inform both the design of CUREs and their placement in a curriculum.

**CURRENT KNOWLEDGE FROM ASSESSMENT OF CUREs**

With this definition in mind, the meeting then turned to summarizing what is known from the study of CUREs, primarily in biology and chemistry. Assessment and evaluation of CUREs has been limited to a handful of multisite programs (e.g., Goodner et al., 2003; Hatfull et al., 2006; Lopatto et al., 2008, Caruso et al., 2009; Shaffer et al., 2010; Harrison et al., 2011) and projects led by individual instructors (e.g., Drew and Triplett 2008; Siritunga et al., 2011). For the most part, these studies have emphasized student perceptions of the
outcomes they realize from participating in course-based research, such as the gains they have made in research skills or clarification of their intentions to pursue further education or careers in science. To date, very few studies of student learning during CUREs have been framed according to learning theories. With a few exceptions, studies of CUREs have not described pathways that students take to arrive at specific outcomes—in other words, what aspects of the CURE are important for students to achieve both short- and long-term gains.

Some studies have compared CURE instruction with research internships and have found, in general, that students report many of the same gains (e.g., Shaffer et al., 2010). A handful of studies have compared student outcomes from CUREs with those from other laboratory learning experiences. For example, Russell and Weaver (2011) compared students’ views of the nature of science after completing a traditional laboratory, an inquiry laboratory, or a CURE. The researchers used an established approach developed by Leidner and colleagues (2002) to assess students’ views of the nature of science, but it is not clear whether students in this study chose to enroll in a traditional or CURE course or whether the groups differed in other ways that might influence the extent to which their views changed following their lab experiences. Students in all three environments—traditional, inquiry, and CURE—made gains in their views of the nature of scientific knowledge as experimental and theory based, but only students in the CURE showed progress in their views of science as creative and process based. When students who participated in a CURE or a traditional lab were queried 2 or 3 yr afterward, they continued to differ in their perceptions of the gains they made in understanding how to do research and in their confidence in doing research (Szteinberg and Weaver, 2013).

In another study, Rowland and colleagues (2012) compared student reports of outcomes from what they called an active-learning laboratory undergraduate research experience (ALLURE, which is similar to a CURE) with those from a traditional lab course. Students could choose the ALLURE or traditional instruction, which may have resulted in a self-selection bias. Students in both environments reported increased confidence in their lab skills, including technical skills (e.g., pipetting) and analytical skills (e.g., deciding whether one experimental approach is better than another). Generally, students reported similar skill gains in both environments, indicating that students can develop confidence in their lab skills during both traditional and CURE/ALLURE experiences.

Most studies reporting assessment of CUREs in the life sciences have made use of the Classroom Undergraduate Research Experiences (CURE) Survey (Lopatto and Tobias, 2010). The CURE Survey comprises three elements: 1) instructor report of the extent to which the learning experience resembles the practice of science research (e.g., the outcomes of the research are unknown, students have some input into the focus or design of the research); 2) student report of learning gains; and 3) student report of attitudes toward science. A series of Likert-type items probe students’ attitudes toward science and their educational and career interests, as well as students’ perceptions of the learning experience, the nature of science, their own learning styles, and the science-related skills they developed from participating in a CURE. Use of the CURE Survey has been an important first step in assessing student outcomes of these kinds of experiences. Yet this instrument is limited as a measure of the nature and outcomes of CUREs because some important information is missing about its overall validity. No information is available about its dimensionality—that is do student responses to survey items meant to represent similar underlying concepts correlate with each other, while correlating less with items meant to represent dissimilar concepts? For example, do responses to items about career interests correlate with themselves highly, but correlate less with items focused on attitudes toward science, a dissimilar concept? Other validity questions are also not addressed. For instance, does the survey measure all important aspects of CUREs and CURE outcomes, or are important variables missing? Is the survey useful for measuring a variety of CUREs in different settings, such as CUREs for majors or nonmajors, or CUREs at an introductory or advanced levels? Finally, is the survey a reliable measure—does the survey measure outcomes consistently over time and across different individuals and settings? To be consistent with the definition of CUREs given above, an assessment instrument must both touch on all five dimensions and elicit responses that capture other important aspects of CURE instruction that may be missing from this description. This will help ensure that the instrument has “content validity” (Trochim, 2006), meaning that the instrument can be used to measure all of the features important in a CURE learning experience.

The CURE Survey relies on student perceptions of their own knowledge and skill gains, and like other such instruments, it is subject to concerns about the validity of self-report of learning gains. There is a very broad range of correlations between self-report measures of learning and measurements such as tests or expert judgments. Depending on which measures are compared, there may be a strong correlation, or almost no correlation, between self-reported data and relevant criteria (Falchikov and Boud, 1989). Validity problems with self-assessment can result from poor survey design, with survey items interpreted differently by different students, or from items designed in such a way that students are unable to recall key information or experiences (Bowman 2011; Porter et al., 2011). The tendency of respondents to give socially desirable answers is a familiar problem with self-reporting. Bowman and Hill (2011) found that student self-reporting of educational outcomes is subject to social bias; students respond more positively because they are either implicitly or explicitly aware of the desired response. A guarantee of anonymity mitigates this validity threat (Albanese et al., 2006). Respondents also give more valid responses when they have a clear idea of what they are assessing and have received frequent and clear feedback about their progress and abilities from others, and when respondents can remember what they did during the assessment period (Kuh, 2001). For example, in her study of the outcomes of undergraduate science research internships, Kardash (2000) compared perceptions of both student interns and faculty mentors of the gains interns made from participating in research. She found good agreement between interns and mentors on some skills, such as understanding concepts in the field and collecting data, but statistically significantly differences between mentor and intern ratings of other skills, with interns rating themselves more positively on their understanding of the importance of controls in research, their abilities to interpret results in light of original hypotheses, and their abilities to relate results to...
the “bigger picture.” More research is needed to understand the extent to which different students (majors, nonmajors, introductory, advanced, etc.) are able to accurately self-assess the diverse knowledge and skills they may develop from participating in CUREs.

A few studies have focused on the psychosocial outcomes of participating in CUREs. One such study, conducted by Hanauer and colleagues (2012), documented the extent to which students developed a sense of ownership of the science projects they completed in a traditional laboratory course, a CURE involving fieldwork, or a research internship. Using linguistic analysis, the authors found that students in the CURE reported a stronger sense of ownership of their research projects compared with students who participated in traditional lab courses and research internships (Hanauer et al., 2012; Hanauer and Dolan, in press, 2014); these students also reported higher levels of persistence in science or in traditional lab courses and research internships (Hanauer et al., 2012). Although the inferred relationship needs to be explored with a larger group of students and a more diverse set of CUREs, these results suggest that it is important to consider ownership and other psychosocial outcomes in future research and evaluation of CUREs.

A few studies have explored whether and how different students experience CUREs differently and, in turn, realize different outcomes from CUREs. This is an especially noteworthy gap in the knowledge base, given the calls to engage all students in research experiences and that research has suggested that different students may realize different outcomes from participating in research (e.g., AAAS, 2011; Thiry et al., 2012). In one such study, Alkaher and Dolan (in press, 2014) interviewed students enrolled in a CURE, the Partnership for Research and Education in Plants for Undergraduates, at three different types of institutions (i.e., community college, liberal arts college, research university) in order to examine whether and how their sense of scientific self-authorship shifted during the CURE. Baxter-Magolda (1992) defined self-authorship as the “internal capacity to define one’s beliefs, relations, and social identity” or, in this context, how one sees oneself with respect to science knowledge—as a consumer, user, or producer. Developing a sense of scientific self-authorship may be an important predictor of persistence in science, as students move from simply consuming science knowledge as it is presented to becoming critical users of knowledge and to seeing themselves as capable of contributing to the scientific body of knowledge. Alkaher and Dolan (in press, 2014) found that some CURE students made progress in their self-authorship because they perceived the CURE goals as important to the scientific community, yet the tasks were within their capacity to make a meaningful contribution. In contrast, other students struggled with the discovery nature of the CURE in comparison with their prior traditional lab learning experiences. They perceived their inability to find the “right answer” as reflecting their inability to do science. More research is needed to determine whether and how students’ backgrounds, motives, and interests influence how they experience CUREs, and whether they realize different outcomes as a result.

NEXT STEPS FOR CURE ASSESSMENT

Our discussion and collective knowledge of research on CUREs and undergraduate research internships revealed several gaps in our understanding of CUREs, which can be addressed by:

- Defining frameworks and learning theories that may help explain how students are influenced by participating in CUREs, and utilizing these frameworks or theories to design and study CUREs;
- Identifying and measuring the full range of important outcomes likely to occur in CURE contexts;
- Using valid and reliable measures, some of which have been used to study research internships or other undergraduate learning experiences and could be adapted for CURE use, as well as developing and testing new tools to assess CUREs specifically (see Weiss and Sosulski [2003] or Trochim [2006] for general explanations of validity and reliability in social science measurement);
- Establishing which outcomes are best documented using self-reporting, and developing new tools or adapting existing tools to measure other outcomes; and
- Gathering empirical evidence to identify the distinctive dimensions of CUREs and ways to characterize the degree to which they are present in a given CURE, as well as conducting investigations to characterize relationships between particular CURE dimensions or activities and student outcomes.

Following these recommendations will require a collective, scholarly effort involving many education researchers and evaluators and many CUREs that are diverse in terms of students, instructors, activities, and institutional contexts. We suggest that priorities of this collective effort should be to:

1. Use current knowledge from the study of CUREs, research internships, and other relevant forms of laboratory instruction (e.g., inquiry) to define short-, medium-, and long-term outcomes that may result from student participation in CUREs;
2. Observe and characterize many diverse CUREs to identify the activities within CUREs likely to directly result in these short-term outcomes, delineating both rewards and difficulties students encounter as they participate;
3. Use frameworks or theories and current knowledge to hypothesize pathways students may take toward achieving long-term outcomes—the connections between activities and short-, medium-, and long-term outcomes;
4. Determine whether one can identify key short- and medium-term outcomes that serve as important “linchpins” or connecting points through which students progress to achieve desired long-term outcomes; and
5. Assess the extent to which students achieve these key outcomes as a result of CURE instruction, using existing or novel instruments (e.g., surveys, interview protocols, tests) that have been demonstrated to be valid and reliable measures of the desired outcomes.

At the front end, this process will require increased application of learning theories and consideration of the supporting research literature, but it is likely to result in many highly testable hypotheses and a more focused and informative approach to CURE assessment overall. For example, if we can define pathways from activities to outcomes, instructors will
be better able to select activities to include or emphasize during CURE instruction and decide which short-term outcomes to assess. Education researchers and evaluators will be better able to hypothesize which aspects of CURE instruction are most critical for desired student outcomes and the most salient to study.

Drawing from many of the references cited in this report, we have drafted a logic model for CURE instruction (Figure 1) as the first step in this process. (For more on logic models, see guidance from the W. K. Kellogg Foundation [2006].) The model includes the range of contexts, activities, outputs, and outcomes of CUREs that arose during our discussion. The model also illustrates hypothetical relationships between time, participation in CUREs, and short- and long-term outcomes resulting from CURE activities.

It is important to recognize that, given the limited time frame and scope of any single CURE, students will not participate in all possible activities or achieve all possible outcomes depicted in the model. Rather, CURE instructors or evaluators could define a particular path and use it as a guide for designing program evaluations and assessing student outcomes. Figure 2 presents an example of how to do this with a focus on a subset of CURE activities and outcomes. It is a simplified pathway model based on findings from the research on undergraduate research internships and CUREs summarized above. Boxes in this model are potentially measurable waypoints, or steps, on a path that connects student participation in three CURE activities with the short-term outcomes students may realize during the CURE, medium-term outcomes they may realize at the end of or after the CURE, and potential long-term outcomes. Although each pathway is supported by evidence or hypotheses from the study of CUREs and research internships, these are not the only means to achieve long-term outcomes, and they do not often act alone. Rather, the model is intended to illustrate that certain short- and medium-term outcomes are likely to have a positive effect on linked long-term outcomes. See Urban and Trochim (2009) for a more detailed discussion of this approach.

We explain below the example depicted in Figure 2, referencing explicit waypoints on the path with italics. This model is grounded in situated-learning theory (Lave and Wenger, 1991), which proposes that learning involves engagement in
a “community of practice,” a group of people working on a common problem or endeavor (e.g., addressing a particular research question) and using a common set of practices (e.g., science practices). Situated-learning theory envisions learning as doing (e.g., presenting and evaluating work) and as belonging (e.g., interacting with faculty and peers, building networks), factors integral to becoming a practitioner (Wenger, 2008)—in the case of CUREs, becoming a scientist.

Retention in a science major is a desired and measurable long-term outcome (bottom of Figure 2) that indicates students are making progress in becoming scientists and has been shown to result from participation in research (Perna et al., 2009; Eagan et al., 2013). Based on situated-learning theory, we hypothesize that three activities students might engage in are likely to lead to retention in a science major: design methods, present their work, and evaluate their own and others’ work.
during their research experience (Caruso et al., 2009; Harrison et al., 2011; Hanauer et al., 2012). These activities reflect the dimensions of “use of scientific practices” and “collaboration” described above.

Following the right-hand path in the model, when students present their work and evaluate their own and others’ work, they will likely interact with each other and with faculty (Eagan et al., 2011). Interactions with faculty and interactions with peers may lead to improvements in students’ communication and collaboration skills, including their abilities to defend their work, negotiate, and make decisions about their research based on interactions (Ryder et al., 1999; Alexander et al., 2000; Seymour et al., 2004). Through these interactions, students may expand their professional networks, which may in turn offer increased access to mentoring (Packard, 2004; Eagan et al., 2011). Mentoring relationships, especially with faculty, connect undergraduates to networks that promote their education and career development by building their sense of scientific identity and defining their role within the broader scientific community (Crisp and Cruz, 2009; Hanauer, 2010; Thiry et al., 2010; Thiry and Laursen, 2011; Stanton-Salazar, 2011). Peer and faculty relationships also offer socio-emotional support that can foster students’ resilience and their ability to navigate the uncertainty inherent to science research (Chemers et al., 2011; Thiry and Laursen, 2011). Finally, research on factors that lead to retention in science majors indicates that increased science identity (Laursen et al., 2010; Estrada et al., 2011), ability to navigate uncertainty, and resilience are important precursors to a sense of belonging and ultimate retention (Gregerman et al., 1998; Zeldin and Pajares, 2000; Maton and Harabowsky, 2004; Seymour et al., 2004). The model also suggests that access to mentoring is a linchpin, a short- to medium-term outcome that serves as a connecting point through which activities are linked to long-term outcomes. Thus, access to mentoring might be assessed to diagnose students’ progress along the top pathway and predict the likelihood that they will achieve long-term outcomes. (For more insight into why assessing linchpins is particularly informative, see Urban and Trochim [2009].)

Examples of measures that may be useful for testing aspects of this model and for which validity and reliability information is available include: the scientific identity scale developed by Chemers and colleagues (2011) and revised by Estrada and colleagues (2011); the student cohesiveness, teacher support, and cooperation scales of the What Is Happening in This Class? questionnaire (Dorman, 2003); and the faculty mentorship items published by Eagan and colleagues (2011). Data will need to be collected and analyzed using standard validation procedures to determine the usefulness of these scales for studying CUREs. Qualitative data from interviews or focus groups can be used to determine that students perceive these items as measuring relevant aspects of their CURE experiences and to confirm that they are interpreting the questions as intended. For example, developers of the Undergraduate Research Student Self-Assessment instrument used extensive interview data to identify key dimensions of student outcomes from research apprenticeship experiences, and then think-aloud interviews to test and refine the wording of survey items (Hunter et al., 2009). Interviews can also establish whether items apply to different groups of students. For example, items in the scientific identity scale (e.g., “I feel like I belong in the field of science”) may seem relevant, and thus “valid,” to science majors but not to non-science majors. Similarly, the faculty-mentoring items noted above (Eagan et al., 2011) include questions about whether faculty provided, for example, “encouragement to pursue graduate or professional study” or “an opportunity to work on a research project.” The first item will be most relevant to students who are enrolled in an advanced rather than an introductory CURE, while the second may be relevant only to students early enough in their undergraduate careers to have time to pursue a research internship. In addition, students may interpret the phrase “opportunity to work on a research project” in ways that are unrelated to mentorship by faculty, especially in the context of a CURE class with its research focus. Statistical analyses (e.g., factor analysis, calculation of Cronbach’s alpha; Netemeyer et al., 2003) should confirm that the scales are consistent and stable—are they measuring what they are intended to measure and do they do so consistently? Such analyses would help determine whether students are responding as anticipated to particular items or scales and whether instruments developed to measure student outcomes of research internships can detect student growth from participation in CUREs, which are different experiences.

We can also follow the left-hand path in this model with a focus on the CURE activities of designing methods and presenting work. This path is grounded in Baxter Magolda’s (2003) work on students’ epistemological development and her theory of self-authorship. Specifically, as students take ownership of their learning, they transition from seeing themselves as consumers of knowledge to seeing themselves as producers of knowledge. Some students who design their own methods and present their work report an increased sense of ownership of the research (Hanauer et al., 2012; Hanauer and Dolan, 2014). Increased ownership has been shown to improve motivation and self-efficacy. Self-efficacy and motivation work in a positive-feedback loop to enhance one another and contribute to development of long-term outcomes, such as increased resilience (Graham et al., 2013). Social cognitive theory is useful for explaining this relationship: if people believe they are capable of accomplishing a task—described in the literature as self-efficacy—they are more likely to put forth effort, persist in the task, and be resilient in the face of failure (Bandura, 1986; Zeldin and Pajares, 2000). Self-efficacy has also been positively related to science identity (Zeldin and Pajares, 2000; Seymour et al., 2004; Hanauer, 2010; Estrada et al., 2011; Adedokun et al., 2013). Thus, self-efficacy becomes a linchpin that interacts closely with motivation and can be connected to retention in a science major. Existing measures that may be useful for testing this model and for which validity and reliability information is available include: the Project Ownership Survey (Hanauer and Dolan, 2014), scientific self-efficacy and scientific identity scales (Chemers et al., 2011; Estrada et al., 2011); and the self-authorship items from the Career Decision Making Survey (Creamer et al., 2010). Again, data would need to be collected and analyzed using standard validation procedures to determine the usefulness of these scales for studying CUREs.

When considering what to include in a model or which pathways to emphasize, we encourage CURE stakeholders to remember that each CURE is in its own stage of development and has its own life cycle. Some are just starting and others are well established. CUREs at the beginning stages of
implementation are likely to be better served by evaluating how well the program is being implemented before evaluating downstream student outcomes. Thus, early in the development of a CURE, those who are assessing CUREs may want to model a limited set of activities, outputs, and short-term outcomes. CUREs at later stages of development may focus more of their evaluation efforts on long-term student outcomes because earlier evaluations have demonstrated stability of the program's implementation. At this point, findings regarding student outcomes can more readily be attributed to participation in the CURE.

Last, we would like to draw some comparisons between CUREs and research internships because these different experiences are likely to offer unique and complementary ways of engaging undergraduates in research that could be informative for CURE assessment. As noted above, a handful of studies indicate that CURE students may realize some of the same outcomes observed for students in research internships (Goodner et al., 2003; Drew and Triplett 2008; Lopatto et al., 2008; Caruso et al., 2009; Shaffer et al., 2010; Harrison et al., 2011). Yet, differences between CUREs and research internships (Table 1) are likely to influence the extent to which students achieve any particular outcome. For example, CUREs may offer different opportunities for student input and autonomy (Pate et al., 2009; Hanauer et al., 2012; Hanauer and Dolan, 2014, Table 2). The structure of CUREs may allow undergraduates to assume more responsibility in project decision making and take on leadership roles that are less often available in research internships. CUREs may involve more structured group work, providing avenues for students to develop analytical and collaboration skills as they explain or defend their thinking and provide feedback to one another. In addition, CURE students may have increased opportunities to develop and express skepticism because they are less likely to see their peers as authority figures.

Alternatively, some CURE characteristics may limit the nature or extent of outcomes that students realize. CUREs take place in classroom environments with a much higher student–faculty ratio than is typical of UREs. With fewer experienced researchers to model scientific practices and provide feedback, students may be less likely to develop a strong understanding of the nature of science or a scientific identity. The amount of time students may spend doing the work in a CURE course is likely to be significantly less than what they would spend in a research internship. Students who enroll in CURE courses may be less interested in research, which may affect their own and classmates' motivation and longer-term outcomes related to motivation. Research interns are more likely to develop close collegial relationships with faculty and other researchers, such as graduate students, postdoctoral researchers, and other research staff, who can in turn expand their professional network. In addition, CURE instructors may have limited specialized knowledge of the science that underpins the CURE. Thus, CURE students may not have access to sufficient mentorship or expertise to maximize the scientific and learning outcomes.

SUMMARY

This report is a first attempt to capture the distinct characteristics of CUREs and discuss ways in which they can be systematically evaluated. Utilizing current research on CUREs and on research internships, we identify and describe five dimensions of CURE instruction: use of science practices, discovery, broader relevance or importance, iteration, and collaboration. We describe how these elements might vary among different laboratory learning experiences and recommend an approach to CURE assessment that can characterize CURE activities and outcomes. We hope that our discussion draws attention to the importance of developing, observing, and characterizing many diverse CUREs. We also hope that this report successfully highlights the enormous potential of CUREs, not only to support students in becoming scientists, but also to provide research experiences to increasing numbers of students who will enter the workforce as teachers, employers, entrepreneurs, and young professionals. We intend for this report to serve as a starting point for a series of informed discussions and education research projects that will lead to far greater understanding of the usages, value, and impacts of CUREs, ultimately resulting in cost-effective, widely accessible, quality research experiences for a large number of undergraduate students.

ACKNOWLEDGMENTS

Thanks to Sandra Laursen for taking a leadership role in organizing this meeting. Support for the meeting was provided by a grant from the NSF (DBI-1061874). The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of the NSF.

REFERENCES


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


Although researchers in undergraduate science, technology, engineering, and mathematics education are currently using several methods to analyze learning gains from pre- and posttest data, the most commonly used approaches have significant shortcomings. Chief among these is the inability to distinguish whether differences in learning gains are due to the effect of an instructional intervention or to differences in student characteristics when students cannot be assigned to control and treatment groups at random. Using pre- and posttest scores from an introductory biology course, we illustrate how the methods currently in wide use can lead to erroneous conclusions, and how multiple linear regression offers an effective framework for distinguishing the impact of an instructional intervention from the impact of student characteristics on test score gains. In general, we recommend that researchers always use student-level regression models that control for possible differences in student ability and preparation to estimate the effect of any nonrandomized instructional intervention on student performance.

INTRODUCTION

For the past several decades, discipline-based education researchers have focused on testing whether educational interventions in college science classrooms lead to improved student understanding and performance. Most interventions are given at the classroom level, meaning that all students in a given classroom receive the intervention. For example, all students in a class may be exposed to a new multimedia program (Aly et al., 2004), asked to participate in reciprocal peer tutoring (Fantuzzo et al., 1989), or taught in a workshop or studio format (Udovic et al., 2002).

To evaluate the impact of educational interventions like these on student performance, researchers typically collect student test scores before and after the intervention—that is, from a pretest and a posttest. Although some researchers are interested in whether student scores improve after instruction (see Arwood, 2004; McConnell et al., 2006; Nam and Ito, 2011), most are interested in demonstrating that student test scores improve more in treatment classrooms than in control classrooms—that is, in sections that do receive the intervention versus sections that do not.

What is the best way to analyze pre–post data in this setting? At least four different methods for determining whether learning gains differ in the treatment and control classrooms are commonly used in the science, technology, engineering, and mathematics (STEM) education literature: comparing 1) raw change scores (e.g., Udovic et al., 2002); 2) normalized gain scores (Hake, 1998); 3) normalized change scores (Marx and Cummings, 2007); and 4) effect sizes (Andrews et al., 2011). Unfortunately, none of these methods accounts for a fundamental problem: rating for student equivalence, or lack thereof, in the classrooms being compared.

The problem of student nonequivalence is pervasive, because it is seldom possible to use randomization to control for differences in student ability or preparation (but see Fantuzzo et al., 1989; Buzzell et al., 2002; Aly et al., 2004; Bilgin et al., 2009). While nonrandomized designs are often unavoidable—it is very difficult to convince a registrar's
office to randomly assign students to courses—they raise difficult questions about interpreting results. Namely, researchers who use the methods listed above have no way of knowing whether observed differences in learning gains between the treatment and control classes are due to the impact of the intervention itself or to differences between treatment and control classes—including the instructor, the instructional techniques used, and student characteristics—that are completely independent of the intervention.

In this paper, we use test score data from two sections of a college-level introductory biology course to illustrate how each of the four commonly used methods can lead to misleading conclusions. The two sections were taught by the same instructor, in the same term, using identical instructional techniques. However, due to a scheduling conflict during that term, the students enrolled in one of the sections had substantially better academic qualifications, on average, than students in the other section. We show that each of the four methods commonly used to assess educational interventions in college STEM classrooms would support the conclusion that an “instructional intervention” in the higher-performing section led to larger student learning gains, when in fact there was no intervention at all.

We propose a solution to the problem by introducing an approach that is ubiquitous in many other research areas but currently underused in the STEM education literature: multiple linear regression analysis. Specifically, we employ a student-level regression model that controls for observable differences between students in the treatment and control classes and demonstrate that it leads to the correct conclusion: differences in learning gains between the two sections are driven by differences in the composition of the students in the two sections, not by any intervention that was given in one section or the other. We argue that to estimate an unbiased intervention effect when analyzing data from non-random experimental designs, researchers must account for student background in a regression framework.

REVIEWING EXISTING METHODS FOR ANALYZING PRE–POST DATA

Before introducing regression approaches for analyzing pre–post data, we provide a brief review of the four approaches commonly used in the undergraduate STEM literature to analyze pre/posttest data and discuss some relative advantages and disadvantages of each. However, we stress that none of these four methods accounts for possible differences in the student composition of the treatment and control courses.

Raw Change Scores

Udovic et al. (2002) compare student learning gains in a “workshop” introductory biology course, which included numerous active-learning activities, with learning gains in comparison courses taught primarily through lectures. Like Dori et al. (2007), Fallahi (2008), and Linsey et al. (2007, 2009), Udovic and colleagues use a t test to compare what we refer to as “raw change scores” between treatment and control classes. Raw change scores are simply the difference between the postscore and the prescore. Udovic and coworkers concluded that if student scores in the treatment course improve more, on average, from the pretest to the posttest than do student scores in the control course, then the gains must be due to the intervention in the treatment courses. This procedure is identical—meaning that it will result in the same p value and conclusions regarding the effect of the intervention—to the two-way repeated-measures analysis of variance (ANOVA) used by Martin et al. (2007).

In both the treatment and control classes, the authors compute the average raw change for each of the 11 questions on their pre- and posttest. The mean raw change was higher in the treatment classes than in the control classes for all 11 questions, and the t test rejected the null hypothesis that the mean raw change scores were equal for seven of the 11 questions. The authors conclude that the active-learning strategies in the treatment courses had a significant impact on student learning gains.

Analyzing raw change is attractive in terms of simplicity but does not account for the observation that students with low scores on a pretest have more to gain than students who score higher. The problem arises because test scores are bounded—meaning that they have an upper limit. To account for differences in “ease of improving” from pre to post, researchers have used two methods for standardizing or normalizing gain scores, one at the classroom level and one at the student level.

Normalized Gain Scores

Hake (1998) compares student learning gains on the Force Concept Inventory across 62 different introductory physics courses. In 48 of these courses, instructors had made substantial use of interactive-engagement methods. Hake considers these the treatment classes, while the 14 courses that were based on traditional lecturing are the control classes. For each class, Hake calculates the “average normalized gain,” symbolized <g>, as the ratio of the average gain from pretest to posttest to the maximum possible gain (⟨g⟩ = \frac{<post−pre>}{100−<pre>}), where pre- and postscores are expressed as the average percent correct in each class in the study. He reports that the average normalized gain in the treatment courses was 0.48 ± 0.14 SD, while the average normalized gain in the control courses was 0.23 ± 0.04 SD. Although he did not perform a formal statistical hypothesis test, he concludes that interactive-engagement methods have a significant positive impact on student learning gains. Had he performed a t test of the average normalized gains, the p value would have been <0.001—more than enough evidence to make the same conclusion.

Reporting <g>, the normalized learning gain, became popular in the undergraduate STEM education literature for several reasons. First, by normalizing by the maximum gain possible in each class, it accounts for the fact that some classrooms have more room to gain than other classrooms. A class that scores 80% on the pretest and a 90% on the posttest has an average normalized gain of 0.5, matching a class that scores 60% on the pretest and 80% on the posttest (i.e., each class gained exactly half of the amount it could have gained on the posttest). Second, the size of Hake’s initial study made it possible for researchers to compare learning gains informally across classrooms, even if their own study did not include enough classes to make a formal statistical test possible. That
Within classrooms rather than across classrooms (see Smith et al., 2009). Finally, in studies with large numbers of treatment and control classrooms, \( g \) can be used to formally test whether learning gains in the treatment classes are larger than in the control classes (e.g., Redish and Steinberg, 1999; LoPresto and Murrell, 2009). However, using \( g \) results in low sample sizes and thus poor statistical power, because it uses the class as the unit of analysis instead of using individual students.

**Normalized Change Scores**

Marx and Cummings (2007) created a student-level alternative to Hake’s normalized gain measure called the normalized change score, symbolized \( c \). Instead of computing learning gains at the classroom level with \( g \), Marx and Cummings advocate calculating learning gains at the student level, using the following formula:

\[
\begin{align*}
    c &= \begin{cases} 
        \text{post} - \text{pre} & \text{if post} > \text{pre} \\
        100 - \text{pre} & \text{drop if post} = \text{pre} = 100 \text{ or 0} \\
        0 & \text{if 0 < post} = \text{pre} < 100 \\
        \text{post} - \text{pre} & \text{if post} < \text{pre} \\
    \end{cases}
\end{align*}
\]

For students who score higher on the posttest than the pretest, the student-level normalized change score is computed similarly to a classroom-level normalized gain. The last three possibilities deal with unusual circumstances: students who score 0 or 100 on both the pre- and posttest are dropped; students who score the same on the pre- and posttest get a 0; and students who score lower on the posttest than the pretest have this negative gain scaled by the possible number of points they could have lost.

Because normalized change scores compare learning gains for students rather than for classrooms, they have two substantial advantages over normalized gain scores. First, they can be used to compare the impact of interventions assigned within classrooms rather than across classrooms (see Smith et al., 2011, for an example). Second, because the observations are at the student level rather than at the classroom level, the sample size is substantially larger compared with using normalized change scores, providing increased statistical power.

Normalized change scores have an important limitation, however. If students get a perfect score on the posttest, their \( c \) is 1 no matter whether their prescore was 1% or 99%. Similarly, if students score the same on the pre- and posttest, their score is 0, no matter whether their prescore was 1% or 99%. In these cases, the goal of normalizing for “ease of improvement” is lost.

**Effect Sizes**

Andrews and colleagues (2011) collect pre/posttest score data on the conceptual inventory of natural selection from a sample of introductory biology courses around the United States, and compare learning gains from courses in which instructors used different numbers of active-learning exercises per week.

To quantify learning gains at the classroom level, they use a metric known as an effect size. Effect sizes are commonly used in meta-analyses because they put estimated treatment effects from different studies in a common scale. For example, researchers can calculate a standardized mean difference, which expresses the difference between groups in units of SD, using Cohen’s \( d \) statistic or a variant called Hedges’ \( g \). With pre–post data from identical assessments, it is appropriate to use a modification of Cohen’s \( d \) that accounts for the same students being tested twice (see Becker, 1988; Dunlap et al., 1996). Thus, Andrews et al. calculate the effect size for each class as:

\[
d = \frac{X_{\text{post}} - X_{\text{pre}}}{s_d \sqrt{2n}}
\]

where \( X_{\text{post}} \) and \( X_{\text{pre}} \) are the average scores on the post- and pretests, \( s_d \) is the SD of the raw gain scores, and \( r \) is the correlation between student scores on the pre- and posttests. Andrews and colleagues estimate a classroom-level linear regression using each class’s effect size as the dependent variable, and find that the number of active-learning exercises used per week has no relationship to student learning gains.

Andrews and colleagues’ (2011) use of linear regression is an important addition to the undergraduate STEM education literature, as it allows them to control for factors other than active learning—such as the instructor’s position and years of teaching experience, class size, and student-rated course difficulty—that could influence learning gains in the treatment and control classrooms. However, Andrews and colleagues estimate their regression at the classroom level and do not have access to student characteristics that can be used as control variables. A large K–12 literature (e.g., Rockoff 2004; Rivkin et al., 2005) demonstrates that observable student characteristics are often correlated not just with student performance but also with student learning gains. Thus, this approach—like the prior three methods we reviewed—does not account for the possibility that differences in student learning gains, or lack thereof, are due to differences in the characteristics of students in the treatment and control classrooms rather than to the effect of the intervention.

**CONTROLLING FOR STUDENT NONEQUIVALENCE: THE PROBLEM**

To illustrate the importance of controlling for observable student characteristics in the treatment and control classes when evaluating the impact of nonrandomized educational interventions, we apply the four methods above to pre- and posttest scores from two sections of an introductory biology course offered during the Summer of 2012 at the University of Washington. Each section was taught by the same instructor using the exact same materials and instructional strategies. Thus, without knowing anything about the student composition of the two classes, there is no a priori reason to expect different student performance in the two classes. Given that there is actually no treatment at all, this should be an example of a statistical test wherein the null hypothesis—that the treatment had no impact on student learning gains—should not be rejected. We will label one of these sections as the treatment (section A) and the other section as the control (section B).

We will demonstrate that each of the methods above does lead to the conclusion that, for this particular pair of sections
in this particular course, learning gains in the treatment class are higher than in the control class. This would ordinarily be taken as evidence that the “instructional intervention” in the treatment class had a significant impact on student learning gains. But given that there was no intervention at all, we explore whether the student composition of these two particular sections may have contributed to the incorrect conclusion. Throughout the analysis that follows, we interpret the results of all tests of statistical significance at the 90% confidence level—meaning that the \( p \) value must be < 0.1 to reject the null hypothesis. We caution, however, against overreliance on conventional levels of statistical significance.

**Data Overview**

At the start of the term, students in each section took a diagnostic test (Shi et al., 2010), converted to a 100-point scale, that was intended to measure their prior knowledge about the topics to be covered in the course. Then 2 wk into the course, students took an in-class exam on the same material—also graded on a 100-point scale—that covered material taught in the first 2 wk of the course. We will treat the diagnostic test as the pretest and the in-class exam as the posttest. The average pretest scores were 59.8 (SD = 18.1) in section A (the “treatment” section) and 59.3 (SD = 17.0) in section B (the “control” section), and are not significantly different between the two sections (the \( p \) value from a two-sample t test is 0.865). This is important because many authors (e.g., McDaniel et al., 2007) assume that the treatment and control classes have similar incoming characteristics if the pretest scores are not significantly different. The average posttest scores were 72.0 (SD = 15.8) in the treatment section and 67.0 (SD = 15.0) in the control section, which a t test indicates is significantly different (\( p = 0.050 \)). We now analyze these data using the four methods discussed in the preceding section.

**Comparison of Raw Change Scores**

In the treatment class, the average raw change score is 72.0 – 59.8 = 12.2 (SD = 15.0), while in the control class, the average raw change score is 67.0 – 59.3 = 7.7 (SD = 15.8). A t test of the null hypothesis that these average raw change scores are the same gives a \( p \) value of 0.077, which is statistically significant at the 90% confidence level. Thus, with this methodology, there is sufficient evidence to reject the null hypothesis and conclude that student learning gains were greater in the treatment class than in the control class.

**Normalized Gain Scores**

The normalized gain score in the treatment class is \( \frac{72.0 - 59.8}{100 - 59.8} = 0.30 \), while the normalized gain in the control class is \( \frac{72.0 - 59.8}{100 - 59.8} = 0.19 \). Because we are limited to only one treatment class and one control class, there is no way to statistically test whether these normalized gain scores are significantly different. That said, the magnitude of the difference may lead to the conclusion that learning gains were greater in the treatment class than in the control class.

**Normalized Change Scores**

The average normalized change score in the treatment class is 0.31 (SD = 0.29), while the average normalized change score in the control class is 0.19 (SD = 0.29). A t test of the null hypothesis that these average normalized changes scores are the same gives a \( p \) value of 0.012, which is statistically significant. Thus, with this methodology, there is sufficient evidence to reject the null hypothesis and conclude that student learning gains were greater in the treatment class than in the control class.

**Effect Sizes**

The correlation between student scores on the pre- and posttests is \( r = 0.56 \). Thus, the effect size for the treatment class is \( \frac{12.2}{\sqrt{15.0} \cdot \sqrt{101 - 0.56}} = 0.76 \), while the effect size for the control class is \( \frac{7.7}{\sqrt{15.8} \cdot \sqrt{101 - 0.56}} = 0.46 \). As with normalized gains, there is no way to test whether these effect sizes are significantly different with only one treatment and one control class. That said, the magnitude of the difference between the two classes may lead to the conclusion that learning gains in the treatment class were larger than learning gains in the control class.

**Potential Explanation**

Each of the above methods could lead to the conclusion that the intervention in the treatment class had a significant positive impact on student learning gains. But given that there was no intervention at all, there must be another explanation for the observed difference in learning gains. One possibility is that the differences occurred by chance. The \( p \) value for a t test comparing normalized change scores, for example, tells us that there is a 1.2% chance of observing differences this extreme by chance alone. Another more probable explanation, though, is that the student composition of the two sections is driving the differences.

To investigate this hypothesis, we collected data on two measures that should reflect student ability and preparation: incoming undergraduate grade point average (GPA) and final grade in the preceding course in the introductory biology sequence. Due to a scheduling conflict during this particular term, the two sections had substantially different incoming performance levels. Specifically, the average incoming GPA in the treatment class was 3.33 (SD = 0.42), which a t test shows is significantly higher (\( p < 0.001 \)) than the average incoming GPA in the control class, 3.04 (SD = 0.43). Likewise, the prior biology grade averaged 3.09 (SD = 0.76) in the treatment class, which a t test indicates is significantly higher (\( p < 0.001 \)) than the average prior biology grade in the control class, 2.69 (SD = 0.66).

This observation underlines the central message of this paper. The gold standard for evaluating the impact of treatments of any kind—educational or otherwise—is a large randomized controlled trial. If sample sizes are large and if treatments are randomly assigned to the experimental subjects (or students, in this case), then there is no reason to expect the treatment and control groups to differ in any way, except that the treatment group received the treatment, while the control group did not. But in the context of evaluating the impact of interventions in undergraduate STEM classrooms, it is often not feasible to randomly assign students to treatment and control classes. The nonrandomized design that results opens up the possibility that the treatment and control classes will be substantially different, as our example shows. If these differences are correlated with student learning gains, then
any of the methods above runs the risk of attributing observed differences to the impact of the treatment, when in reality they are due to differences in the composition of the groups being compared. This is true even if the treatment had no effect at all.

Given that many interventions in undergraduate STEM education cannot be randomized, is there a way to distinguish the impact of incoming student characteristics from the impact of the intervention itself? We argue that the answer is often yes, and that multiple linear regression can be a useful tool in any such analysis. We introduce this methodology in the next section, apply it to our data, and demonstrate that it leads to the correct conclusion: controlling for incoming student characteristics, there is no statistically significant difference in learning gains between the treatment and control classes in our example. Our goal in the following section is not to provide a rigorous theory of linear regression, but rather to motivate its use for evaluating the impact of educational interventions on student learning gains. We refer interested readers to chapters 3 and 4 of Gelman and Hill (2007) for an accessible discussion of broader considerations in linear regression.

**CONTROLLING FOR STUDENT NONEQUIVALENCE: A SOLUTION**

It is intuitive to think of a student’s performance on a test as a function of many factors: the student’s prior knowledge about the specific topics on the test, the student’s understanding of the larger discipline, the student’s work habits and study skills, and the intervention itself. A linear regression model formalizes this intuition by assuming that an outcome or dependent variable (in this case, a student’s score on the posttest) is a linear function of explanatory (or control) variables and the intervention itself. Linear regression is not the only methodology that allows for this framework, but we will restrict our attention to it for simplicity.

A key step in a linear regression analysis is collecting data about control variables—measurements that can serve as proxies for factors that may influence the outcome variable, other than the treatment of interest. In a pre- and posttest setting, each student’s score on the pretest is one obvious control variable, as the pretest controls for each student’s prior knowledge about the specific topics on the test. In our example, we also collected data on each student’s undergraduate GPA and grade in a previous biology class. The latter may be a reasonable proxy for each student’s understanding of the broader field of biology, while both measures provide some information about each student’s work habits and study skills.

Undergraduate GPA and previous biology grade are certainly not the only variables we could select to control for variation in student preparation and ability. In fact, researchers often have access to more student-level variables than are practical to use. Procedures like stepwise regression can assist researchers in selecting control variables that are most predictive of the outcome variable (see Freeman et al., 2007, for an example). Researchers can also use professional judgment—based on the available variables, data from similar studies in the literature, and their own experience—in selecting control variables. We chose a measure of overall academic performance (undergraduate GPA) and a measure of performance specific to biology (previous biology grade). Although the correlation between these variables is high ($r = 0.84$), we chose both to account for the possibility that they capture different dimensions of student academic background—a decision that is borne out in the results in the next section. We also note that if we had data on many classrooms taught at different times by different instructors, we would also consider controlling for indicators such as time-of-day and class instructor, if the data suggested these indicators were relevant to the outcome being measured.

With these considerations in mind, we define the following variables for each student: $X_{\text{post}}$ is the student’s score on the posttest; $X_{\text{pre}}$ is the student’s score on the pretest; GPA is the student’s undergraduate GPA; Grade is the student’s grade in the introductory course; and Treatment is an indicator for whether the student was in the class that received the intervention (Treatment = 1 if the student is in the treatment class, Treatment = 0 if the student is in the control class). We recommend centering each of the control variables (in this case, Pre, GPA, and Grade) by subtracting the mean of each variable from each student’s value (see Gelman and Hill, 2007, sections 4.1 and 4.2). One possible linear regression model that uses these variables is the following:

$$
X_{\text{post}} = \beta_0 + \beta_1 \times X_{\text{pre}} + \beta_2 \times \text{GPA} + \beta_3 \times \text{Grade} + \beta_4 \times \text{Treatment} + \epsilon \quad (1)
$$

The $\beta$s in Eq. 1 are regression coefficients that describe the relationship between each variable and the student’s postscore:

- $\beta_0$ is the intercept, or the expected postscore for a student with an average prescore, GPA, and prior grade, who did not receive the treatment (note that if the control variables are not centered, $\beta_0$ is the expected postscore for a student with prescore of 0, GPA of 0, etc., which are not meaningful values);
- $\beta_1$ is the expected increase in the postscore for each additional point on the student’s prescore;
- $\beta_2$ is the expected increase in the postscore for each additional GPA point;
- $\beta_3$ is the expected increase in the postscore for each additional grade point from the previous course; and
- $\beta_4$ is the expected increase in the postscore for students who received the intervention relative to students who did not receive the intervention.

In contrast to approaches like type I ANOVA that estimate the effect of each variable sequentially, a linear regression estimates each of these coefficients simultaneously. Thus, each of these regression coefficients should be interpreted as “all else changing one variable while holding all the other variables constant.” The error term $\epsilon$ captures the reality that the regression equation does not perfectly predict each student’s postscore.

Statistical software packages provide an estimate for each regression coefficient and the $p$ value from the $t$ test of the null hypothesis that the coefficient equals zero. For example, consider the coefficient of interest in Eq. 1: $\beta_4$, or the "treatment effect" for students who received the intervention relative to students who did not receive the intervention.
Table 1. Estimated regression coefficients from linear regression Eq. 1

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Estimate</th>
<th>SE</th>
<th>p Value from t testa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept ($\hat{\beta}_0$)</td>
<td>69.61</td>
<td>1.31</td>
<td>&lt;0.0001***</td>
</tr>
<tr>
<td>Prescore ($\hat{\beta}_1$)</td>
<td>0.28</td>
<td>0.06</td>
<td>&lt;0.0001***</td>
</tr>
<tr>
<td>GPA ($\hat{\beta}_2$)</td>
<td>12.31</td>
<td>3.82</td>
<td>0.0016**</td>
</tr>
<tr>
<td>Grade ($\hat{\beta}_3$)</td>
<td>4.37</td>
<td>2.30</td>
<td>0.0995*</td>
</tr>
<tr>
<td>Treatment ($\hat{\beta}_4$)</td>
<td>−0.42</td>
<td>1.91</td>
<td>0.8272</td>
</tr>
</tbody>
</table>

aSignificance levels from two-sided t test: *, p < 0.1; *, p < 0.05; **, p < 0.01; ***, p < 0.001.

The null hypothesis is that this coefficient equals zero—that is, the intervention had no effect. Linear regression provides both an estimate of this treatment effect and a test of whether the treatment effect really is significantly different from zero, controlling for the influence of each of the other variables in the model. Note that if differences in learning gains between the treatment and control classes can be explained by the control variables and not by the intervention itself, then the treatment effect should not be significantly different from zero. On the other hand, if the intervention does have a significant impact on student performance, the null hypothesis should be rejected, and (if the regression model is correctly specified) the estimated treatment effect should quantify the average effect of the intervention on student test scores.

Linear regression makes some important assumptions. While it is beyond the scope of this paper to discuss all of them in depth (see Gelman and Hill, 2007, section 3.6, for more details), there are a few that are particularly important for the present application. The first is that the error term $\epsilon$ is normally distributed. This assumption can be problematic if the maximum score on the test creates a “ceiling effect” that artificially limits the scores of the best students in the class. In this situation, these students will consistently score lower than the model predicts, because there is a violation of the normality assumption. Another assumption is that the influence of the control variables truly is linear. There is no compelling reason, other than mathematical convenience, to assume that the influence of a student’s prescore, GPA, and prior grades on his or her postscore is truly additive as opposed to multiplicative or otherwise nonlinear.

These assumptions are important, and there are many methods to test and relax them (see Gelman and Hill, 2007, chapters 3–6). Here, however, we focus on standard linear regression.

DATA ANALYSIS USING MULTIPLE LINEAR REGRESSION

We now return to the introductory biology data and illustrate that linear regression leads to the correct conclusion that there is no evidence that the “intervention” has a significant impact on student performance, controlling for other differences between students in the treatment and control classes. We estimate the linear regression equation specified in Eq. 1, and report the estimated coefficients $\hat{\beta}_i$ in Table 1. (Note that the “hat” over each regression coefficient indicates that it is an estimate of the parameter $\beta_i$ in Eq. 1.)

These estimates can be interpreted as follows.

- The estimated intercept ($\hat{\beta}_0 = 69.61$) means that the expected score on the posttest for a student with an average prescore, average GPA, and average prior grade, who did not receive the intervention is 69.61.
- The coefficient on the prescore ($\hat{\beta}_1 = 0.28$) means that we expect a student’s score on the posttest to increase by 0.28 points for each additional point the student scores on the pretest, all else equal (i.e., holding all the other variables constant).
- The coefficient on GPA ($\hat{\beta}_2 = 12.31$) means that we expect a student’s score on the posttest to increase by 12.31 points for each additional point in the student’s GPA, all else equal.
- Finally, the coefficient on the prior grade ($\hat{\beta}_3 = 4.37$) means that we expect a student’s score on the posttest to increase by 4.37 points for each additional point in the student’s grade from the previous biology course, all else equal.

For each coefficient, the null hypothesis that the coefficient equals zero is rejected at the 90% confidence level, so we have sufficient evidence to conclude that each of these control variables has an independent, significant correlation with student performance on the posttest. This is extremely important, as it means that even when controlling for a student’s score on the pretest, a student’s GPA and prior grades are still predictive of his or her score on the posttest. This may be due to students with higher GPAs having better study skills and work habits, and therefore preparing more effectively for the posttest. As in many studies, the posttest in our example was awarded course points, while the pretest was not—a situation that may increase the impact of differences in motivation or preparation. Alternatively, it is possible that students who received a better grade in the previous biology course have a better understanding of the broader discipline, which helped them prepare for and answer questions on the posttest.

Finally, the $p$ value of the estimate $\hat{\beta}_4$ from the $t$ test of the null hypothesis of no treatment effect—is 0.827, which means that there is not nearly enough evidence to reject the null hypothesis that the “intervention” is significantly correlated with student performance, controlling for the influence of other student-level characteristics. Given that there was no intervention in the treatment class at all, it is reassuring that the linear regression model leads to this conclusion. This reinforces our central message: It is essential to control for potential student nonequivalence between the treatment and control classes.
control groups when evaluating the impact of a nonrandomized educational intervention.

TOWARD INTEGRATION OF LINEAR REGRESSION IN UNDERGRADUATE STEM RESEARCH

We have shown that existing methods of evaluating interventions in college science classrooms can lead to erroneous conclusions when the interventions are not randomly assigned to students, and that linear regression can help mitigate this problem by controlling for observable characteristics that are also correlated with student learning gains. We caution that estimates from a linear regression do not justify a causal interpretation, except under strict assumptions, and that randomization of the intervention is still the best way to establish a treatment effect.

In our motivating example, we have used a multiple linear regression model to illustrate the simplicity and utility of a regression framework. However, there are many other reasons that education researchers should be drawn to this framework. Although we choose not to control for gender and ethnicity in our regression model, regression can also be used to test whether women, minorities, or any other affinity group are gaining more or less in our classrooms, all else equal. Regression models can also include interaction terms that test whether the intervention has a differential impact on different types of students. Researchers who currently use normalized change scores can simply use these values as the outcome variable in a linear regression. (When doing so, though, we recommend not including prescore as a predictor variable, as prescore is already included in normalized change.) Finally, while it is beyond the scope of this paper to discuss more complex regression methods, an even more rigorous approach could use generalized linear models to model non-linear relationships between student characteristics and test scores, analyze student responses at the individual-question level, or produce unbiased estimates in the presence of a ceiling effect.

The undergraduate STEM education literature has made remarkable strides in recent years, but the methods commonly used to estimate the impact of instructional interventions lead to troubling questions about whether these treatment effects really are due to the interventions. It is possible that none of the results in the studies we reviewed would have changed if the researchers had controlled for student characteristics in a regression framework, but we hope we have illustrated that linear regression should be a component of any analysis of a nonrandomized instructional intervention. It is time for this growing literature to take the next step and ensure that reported treatment effects are the result of the intervention itself, not the students.

ACKNOWLEDGMENTS

We thank Ben Wiggins for discussion and for access to the pre–post testing data on introductory biology students, and Sara Brownell, Sarah Eddy, Eli Theobald, Mary Pat Wenderoth, and two anonymous reviewers for comments that improved the manuscript. The study was funded by a grant from the National Science Foundation (DUE-0942215) and conducted under Human Subjects Division application #44438.

REFERENCES


Essay

Feedback about Teaching in Higher Ed: Neglected Opportunities to Promote Change

Cara Gormally,* Mara Evans,† and Peggy Brickman‡

*Department of Science, Technology, and Mathematics, Gallaudet University, Washington, DC 20002; †Biology Academic Success Center, University of California, Davis, Davis, CA 95616; ‡Department of Plant Biology, University of Georgia, Athens, GA 30602

Submitted December 6, 2013; Revised March 27, 2014; Accepted March 27, 2014

Monitoring Editor: Diane K. O’Dowd

Despite ongoing dissemination of evidence-based teaching strategies, science teaching at the university level is less than reformed. Most college biology instructors could benefit from more sustained support in implementing these strategies. One-time workshops raise awareness of evidence-based practices, but faculty members are more likely to make significant changes in their teaching practices when supported by coaching and feedback. Currently, most instructional feedback occurs via student evaluations, which typically lack specific feedback for improvement and focus on teacher-centered practices, or via drop-in classroom observations and peer evaluation by other instructors, which raise issues for promotion, tenure, and evaluation. The goals of this essay are to summarize the best practices for providing instructional feedback, recommend specific strategies for providing feedback, and suggest areas for further research. Missed opportunities for feedback in teaching are highlighted, and the sharing of instructional expertise is encouraged.

INTRODUCTION

Despite heroic dissemination of evidence-based teaching practices and their documented improvement on student learning (Ebert-May et al., 1997; Derting and Ebert-May, 2010; Crouch and Mazur, 2001; Udovic et al., 2002; Knight and Wood, 2005; Freeman et al., 2007), university science faculty members have been slow to adopt these practices. In a national survey of new physics faculty members, 25% reported they had attended teaching workshops (Henderson, 2008) and 87% of these reported knowledge of one or more evidence-based strategies, yet only 50% of those attending report adopting these practices (Henderson and Dancy, 2009). These faculty members identified several impediments to adoption, including inadequate training, misunderstanding of evidence-based teaching practices, and lack of support for implementation (Dancy and Henderson, 2010). Two separate studies have documented misunderstandings about what is involved in evidence-based teaching. Ebert-May and colleagues (2011) identified a significant discrepancy between the degree to which faculty members report using active learning versus levels of active learning observable in video recordings of their classrooms. A multi-institution investigation of introductory biology courses also revealed that self-reported use of active-learning instruction was not associated with student learning gains (Andrews et al., 2011).

Collectively, this work suggests that one-time workshops raise awareness of evidence-based teaching strategies but are not sufficient for faculty to adopt and successfully use these strategies (National Research Council [NRC], 2012).

We propose that learning to teach, like developing other professional skills, requires acquiring knowledge about performing job-related tasks, but it also must involve feedback and mentoring in order to monitor and improve performance (Hattie and Timperley, 2007; Nielsen, 2011; Finkelstein and Fishbach, 2012). However, college teaching is one of the few vocations that requires neither formal training (Golde and Dore, 2001; Tanner and Allen, 2006; Addy and Blanchard, 2010) nor standard processes for evaluation and supervision (Centra, 1993; Weimer and Lenze, 1994; Johnson and Ryan, 2010).

DOI: 10.1187/cbe.13-12-0235

Address correspondence to: Cara Gormally (cara.gormally@gallaudet.edu).

© 2014 C. Gormally et al. CBE—Life Sciences Education © 2014 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.
Providing Feedback on Teaching

THE NEED FOR FEEDBACK ABOUT TEACHING

Institutions are beginning to recognize the need to offer more substantive and formative instructional feedback to faculty (Seldin, 1999; Bernstein, 2008; Huston and Weaver, 2008; Ismail et al., 2012), although few agree on how to provide it (Johnson and Ryan, 2000). Safavi and colleagues (2013) report that 96% of faculty surveyed (n = 237) desire more meaningful instructional feedback. Currently, faculty members receive the majority of their teaching feedback through student evaluations (Keig, 2000; Locher, 2006), with the occasional peer-teaching observation (Seldin, 1999). There are considerable limitations to both feedback mechanisms.

Student evaluations typically focus on gathering data about student perceptions of teacher-centered behaviors such as instructor enthusiasm, clarity of explanations, and provide only limited opportunities for students to comment on the use of learner-centered pedagogies (Murray, 1983; Cashin, 1990; Marsh and Roche, 1993). This may partially explain the decline in student evaluation scores often mentioned by faculty members who incorporate active learning into their courses (Walker et al., 2008; Brickman et al., 2009; White et al., 2010). Items on student evaluations typically focus on student satisfaction and didactic teaching, rather than measuring learning (d’Apollonia and Abrami, 1997; Aleamoni, 1999; Kolitch and Dean, 1999; Kember et al., 2002). Disciplinary and class-size bias have already been noted as a problem in student evaluations: science and mathematics disciplines garner the lowest student evaluation scores (Cashin, 1990; Ramsden, 1991; Aleamoni, 1999); science courses typically have larger enrollments than arts and humanities courses (Cheng, 2011); and student evaluations are lower in larger classes (Aleamoni and Hexner, 1980; McKeachie, 1990; Franklin, 1991).

Faculty members express reservations about the use of student evaluations, particularly for personnel and tenure decisions, and even opposed them outright when they were first introduced (Hills, 1974; Chandler, 1978; Vasta and Sarmiento, 1979; Dowell and Neal, 1982; Menefee, 1983; Zoller, 1992; Goldman, 1993). Faculty members contend that student evaluations lead to lower morale and job satisfaction and may even motivate faculty to reduce standards on examinations and assignments in an effort to placate students, due to their focus on students’ satisfaction (Ryan, 1980; Schneider, 2013). Faculty members have also expressed concern over the appropriate role for student evaluations of their teaching effectiveness in personnel decisions such as retention, promotion, tenure, and salary increases (Cashin and Downey, 1992).

Others have repeatedly argued that student evaluations improve teaching effectiveness (Overall and Marsh, 1979; Cohen, 1980; Marsh and Roche, 1993). However, as the sole measure of teaching effectiveness or as an impetus to increase active learning in the college classroom, student evaluations are far from adequate. Student evaluations provide few concrete ideas for improving instructional effectiveness or learning outcomes (Cohen and McKeachie, 1980; Abrami et al., 1990) or changing curriculum or course objectives (Neal 1988; Abrami 1989). Instructors find it difficult to reconcile contradictory opinions expressed in student evaluations (Ryan, 1980; Callahan, 1992). Consequently, only a small percentage of faculty members report making changes to their courses as a result of student evaluations (Spencer and Flyr, 1992; Kember et al., 2002; Richardson, 2003). And, as we later discuss in depth, faculty may have little incentive to use the data from student evaluations (Kember et al., 2002; Mervis, 2013). Researchers have documented that pairing student evaluations with qualitative student interviews or peer consultations are much more effective at influencing faculty behavior (Cohen, 1980; Wilson, 1986; Tiberius, 1989; Seldin, 1993). However, these practices are not currently implemented at most universities and are difficult to implement at the scale required by many institutions.

Peer-review approaches for evaluating teaching have also been studied and found lacking (Hutchings, 1995; Quinlan and Bernstein, 1996; Huston and Weaver, 2008). One-time classroom observations conducted by peer faculty typically focus on content accuracy, while offering little input about curricular alignment or objectives (Malik, 1996), and often lack collaboration and support from colleagues (Bernstein, 2008). One-time classroom observations also suffer from additional problems, including but not limited to, faculty lack of expertise in providing instructional feedback (Kremer, 1990), observer bias toward similar teaching style (Centra, 2000), reliability issues and conflicts of interest resulting in reluctance to give a peer negative feedback (Marsh, 1984; Feldman, 1988), and power dynamics requiring delicate maneuvering (Keig and Waggone, 1994). Moreover, one-time observations have been shown to have virtually no impact on faculty teaching, aside from influencing textbook selection (Spencer...
and Flyr, 1992), and may even lead to erroneous inferences (Weimer, 2002). Faculty members are also resistant to the use of summative peer evaluation, which they feel contributes little to tenure and promotion decisions (Iqbal, 2013).

Having considered the purpose of instructional feedback, and current practices, we provide a broad review of the best practices of giving feedback in the next section.

CHARACTERISTICS OF EFFECTIVE FEEDBACK

In general, regardless of the task, feedback is meant to provide advice from a mentor or provider to assist a recipient with modifying and improving future performance. The question is how to best provide feedback so that it results in improved performance of a specific task. There are a host of factors that come into play, from the complexity of the task to the method of imparting feedback to the definition used to judge performance. Although the value of feedback is frequently noted in the literature (Brinko, 1993; Hattie and Timperley, 2007; Ismail et al., 2012), there is little research on what makes feedback given to faculty effective for improving undergraduate teaching (Bernstein, 2008; Stes et al., 2010). For the purposes of this review, we define feedback as “information provided by an agent (e.g., teacher, peer, book, parent, self, experience) regarding aspects of one’s performance or understanding” (Hattie and Timperley, 2007, p. 81). Feedback does not have to be provided by another person; individuals are capable of acquiring feedback through self-reflection. For example, one may learn tasks simply through observing others’ performance (Bandura, 1977; Green and Osborne, 1985). The observer then modifies his or her own behavior by comparisons with others and subsequent self-reflection (Wong, 1985).

We draw on the extensive literature from organizational psychology about the characteristics of feedback that are important for improving workplace performance. For example, Kluger and DeNisi (1996) review the effectiveness of vocational interventions designed to inform recipients about ways to improve their performance on tasks, but exclude feedback related to interpersonal issues. These tasks were as diverse as typing, test performance, and attendance behavior on the job. They caution that feedback does not always result in improved performance and can in fact be detrimental. They conclude that several factors affect the outcome of feedback. These factors include how the task is defined and how feedback is delivered. In work situations, for example, feedback that threatens self-esteem or interferes with the initial stages of learning a new task can have a negative effect on performance (Kluger and DeNisi, 1996). We also draw from literature on the effects of feedback on K–12 student outcomes. Researchers have shown that in testing situations, for example, students do not improve on subsequent tests simply by knowing they missed an item. To improve on subsequent tests, they also need to know the correct answer (Bangert-Drowns, 1991). Finally, we include substantial evidence from the K–12 teacher education literature that immediate and specific instructional feedback supports continuing growth (Brinko, 1993; Scheeler et al., 2004). We also reference the few empirical studies analyzing the effectiveness of feedback, mentoring, and coaching given as part of university faculty instructional development (Stes et al., 2010).

Through review of these and other studies from K–12 teacher education and workplace performance, we identified characteristics of effective feedback (Table 1) that are described in detail below. Effective feedback: 1) clarifies the task by providing instruction and correction; 2) improves motivation that can prompt increased effort; and 3) is perceived as valuable by the recipient, because it is provided by credible sources (Table 1). We propose that feedback about undergraduate teaching that is characterized by these features can lead to tangible benefits, including instructor growth and accolades, increased instructor motivation, and improved student learning.

Effective Feedback Clarifies the Task in a Specific, Timely Manner, with a Consistent Message That Informs Recipients How to Improve

At a fundamental level, feedback provides information useful for measuring performance compared with expectations (a task standard) and provides suggestions to correct discrepancies between one’s performance and that task standard (Hattie and Timperley, 2007). To correct discrepancies, feedback must identify the type and extent of errors and contain suggestions for correcting them (Scheeler et al., 2004). If the task standard the recipient is aiming for is not clear, then feedback is less likely to be effective. For example, physicians in training are able to improve performance when the feedback they receive includes critical incidents that indicate when their performance deviated from the task standard (Wigton et al., 1986). The recipients of this specific feedback understand their evaluations better (Liden et al., 1979). However, if there is no clear task-related standard against which to compare for a novel task, then it should not be surprising that feedback will have little effect. If there are conflicting sources of feedback in the environment (peers, etc.), then the discrepancy may make it difficult to resolve how to integrate the feedback (Kluger and DeNisi, 1996).

Feedback must be concrete and specific: not only is concrete, specific feedback preferred by recipients (Liden and Mitchell, 1985), it is also more effective than general feedback. For example, K–12 teachers are more likely to improve their behaviors (e.g., the amount of time spent asking questions of students or other pacing and prompting behaviors) when they are given specific feedback that includes examples of how to improve rather than just general information, for example, telling them the number of questions they asked students (Englert and Sugai, 1983; Hindman and Polsgrove, 1988; Giebelhaus, 1994; O’Reilly and Renzaglia, 1994).

Feedback has been shown to be most effective when it is provided in a timely manner. In the K–12 setting, researchers compared changes to teaching behaviors following feedback that was delivered immediately or after a delay. Providing feedback after a delay was less effective compared with providing feedback immediately after performance. Immediate feedback involved supervisors interrupting instruction when the teacher incorrectly performed a target behavior, identifying the error for the teacher, asking the teacher how he or she could correct the error, and often providing a more appropriate procedure or modeling the correct behavior (O’Reilly, 1992; O’Reilly and Renzaglia, 1994; Coulter and Grossen, 1997). Similar studies demonstrated that feedback
was more effective at changing teaching behaviors beyond an immediate class session if given over multiple—but not too-frequent occasions (Rezler and Anderson, 1971; Ilgen et al., 1979; Chhokar and Wallin, 1984; Fedor and Buckley, 1987).

Effective feedback provides a consistent message that considers both the recipient’s knowledge and other conflicting messages they may be receiving. Both peers and students explicitly compare teaching performance with that of other instructors (Cavanagh, 1996). McColskey and Leary (1985) refer to this comparative feedback as “norm-referenced.” Norm-referenced feedback that conveyed the message of failure (negative feedback) led to lower self-esteem, expectations, and motivation (McColskey and Leary, 1985). In contrast, “self-referenced” feedback, which compared an individual’s performance with other measures of his or her ability, produced increased feelings of competence, because the feedback attributed the individual’s skills to personal effort and contained higher expectations for future performance (McColskey and Leary, 1985).

One alternative to norm-referenced feedback is Utell’s (2013) facilitative feedback model, which seeks to build skills and expose opportunities for growth. The facilitative feedback model shares similarities with the peer-teaching discussion group model proposed by Anderson et al. (2011). Other models also rely on the establishment of a mentoring relationship between the individuals receiving and providing feedback (Showers, 1984; Centra, 1993; Johnson and Ryan, 2000). In these models, the instructor’s strengths and weaknesses are explicitly identified before a task is performed. During and after performance of the task, the instructor receives feedback from the mentor. The mentor suggests ways for the instructor to improve and highlights areas of strength and future potential. Additionally, meeting before the observation may increase buy-in for the process. This opens the door for two-way conversation, shifting the process from

### Table 1. Providing effective instructional feedback

<table>
<thead>
<tr>
<th>Qualities of effective feedback</th>
<th>Characteristics</th>
<th>Suggestions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Clarifies the task by providing instruction and correction</td>
<td>• Provides instruction</td>
<td>• Teaching and learning conferences</td>
</tr>
<tr>
<td></td>
<td>• Defines a clear standard for how the task should be completed</td>
<td>• Workshops on innovative teaching practices</td>
</tr>
<tr>
<td></td>
<td>• Concrete and specific</td>
<td>• Online video resources</td>
</tr>
<tr>
<td></td>
<td>• Identifies types of errors and provides suggestions for correction</td>
<td>• Feedback is guided by validated classroom observation protocols.</td>
</tr>
<tr>
<td></td>
<td>• Timely (as soon as possible after performance of the task)</td>
<td>• Debrief immediately after the peer observation, rather than months later or at the end of semester.</td>
</tr>
<tr>
<td></td>
<td>• Occurs over multiple occasions</td>
<td>• Observations occur several times during the semester.</td>
</tr>
<tr>
<td></td>
<td>• Consistent, minimizes conflicting messages from students and peers</td>
<td>• Discuss expectations of department and methods for dealing with student resistance.</td>
</tr>
<tr>
<td></td>
<td>• Self-referenced (compared with an individual’s ability and expectations rather than compared with a peer)</td>
<td>• Have a consistent template for peer-teaching evaluations.</td>
</tr>
<tr>
<td></td>
<td>• Does not interfere with the initial stages of learning</td>
<td>• Discuss individual’s concerns and address specific challenges that instructor wishes to solve.</td>
</tr>
<tr>
<td></td>
<td>• Does not threaten self-esteem</td>
<td>• Meet before classroom observation to set up expectations and solicit feedback about specific challenges.</td>
</tr>
<tr>
<td>2. Improves motivation that can prompt increased effort</td>
<td>• Leads to higher goal setting</td>
<td>• Choose a date after the first instructional opportunity.</td>
</tr>
<tr>
<td></td>
<td>• Provides a positive encouraging message</td>
<td>• Highlight areas of strength and areas for improvement as a formative evaluation that is not part of promotion and tenure decisions.</td>
</tr>
<tr>
<td></td>
<td>• Accounts for confidence and experience level</td>
<td>• Focus on student outcomes and changes that result in gains in student achievement.</td>
</tr>
<tr>
<td>3. Perceived as valuable by the recipient because it is provided by a reputable source</td>
<td>• Encourages seeking feedback voluntarily</td>
<td>• Unit head implements peer-coaching model with volunteers.</td>
</tr>
<tr>
<td></td>
<td>• Increases perception of value of feedback to improve job status</td>
<td>• Unit head provides rewards for seeking feedback in the same way he or she rewards positive student evaluations in evaluating faculty performance.</td>
</tr>
<tr>
<td></td>
<td>• Protects the ego and others’ impressions</td>
<td>• Private and developmental rather than public and evaluative. Copies of any written materials provided to the department mention that peer evaluation occurred, not the substance of the discussions.</td>
</tr>
<tr>
<td></td>
<td>• Respected status of feedback provider</td>
<td>• Knowledgeable source of higher status who expresses they are providing feedback for the well-being and improvement of the recipient and for improved student outcomes.</td>
</tr>
</tbody>
</table>
evaluation to coaching, and provides opportunities for the instructor to suggest areas of concern or interest to the mentor (Skinner and Welch, 1996). This type of model accounts for individual differences in experience and presents a consistent message. This could help instructors navigate the conflicting, and frequently negative feedback given by disparate sources.

Effective Feedback Encourages the Instructor, Improving Motivation and Stimulating Increased Effort

Both the tone of feedback and the context in which it is given have both been shown to be important for determining effectiveness. Thinking about business management author Michael Leboeuf’s quote from his 1985 book *The Greatest Management Principle in the World* (Putnam), “what gets rewarded gets done,” reminds us to consider the factors that motivate someone to want to improve at his or her job. Locke and Latham’s (2006) goal-setting theory suggests that providing feedback per se does not improve motivation or performance, but it will do so if it leads to higher goals being set or greater commitment to existing goals. In a meta-analysis of 33 studies, Locke and Latham (1990) report that the setting of specific, challenging goals, instead of easy or vague goals like “doing your best,” consistently led to better performance.

Feedback should be positively framed but not generically positive. Instructors prefer hearing positive feedback over negative feedback (Jussim et al., 1995). Feedback is more easily recalled when it is accompanied by a positive encouraging message compared with negative messages (Podsakoff and Farh, 1989); and positive feedback is considered more accurate (Podsakoff and Farh, 1989; Jussim et al., 1995). In K–12 settings, researchers have demonstrated that the addition of a positive message to noncorrective feedback (e.g., information on the number of times the teacher exhibited a specific behavior) increases the effectiveness of that feedback as compared with noncorrective feedback alone (Cossairt et al., 1973). However, perpetually receiving only positive feedback leads to complacency (Podsakoff and Farh, 1989); perhaps an instructor begins to think, “I am doing so well, I don’t need to improve.”

Feedback providers should consider the confidence and experience of the recipient when choosing the appropriate amount of encouragement. Individuals with lower self-confidence tend to view negative feedback as more accurate (Jussim et al., 1995) and to rely on feedback from external sources rather than from themselves (Ilgen et al., 1979). Novices generally have lower self-esteem, and they indicate a preference for positive feedback. For example, novice learners preferred language instructors who emphasized what students were doing well in the classroom rather than correcting mistakes (Finkelstein and Fishbach, 2012). Experts, however, will seek out negative feedback, indicating more interest in learning what they did wrong and how to correct it (Finkelstein and Fishbach, 2012).

Unfortunately, the common practices for imparting instructional feedback in higher education do not account for differences in instructor self-confidence and experience. Faculty commonly receive negative, or what Utell (2013) refers to as “failure-based feedback,” which focuses on fault-finding in task performance. Failure-based feedback can be found in the two most common types of teaching feedback. Students’ references to evaluations as an opportunity to “vent” (Marlin, 1987; Lindahl and Unger, 2010) or as a “plot to get back at an instructor” (Jacobs, 1987) are examples of fault-finding feedback. Students can also express failure-based feedback by choosing not to enroll in courses, and this feedback can have devastating consequences. For example, one study documented the termination of a faculty member following rising student attrition rates in a course utilizing evidence-based teaching practices (Silverthorn et al., 2006).

An instructor may be less likely to take risks and therefore choose not to adopt evidence-based teaching strategies if these are perceived as too risky or likely to result in failure-based feedback from students or peers.

Feedback Is More Likely to Be Sought If the Potential Benefit Outweighs the Costs

As we reviewed in the Introduction, the current models for receiving feedback in higher education—end-of-course student ratings and peer reviews—are intended to assess competence using a standardized instrument, are prescribed rather than voluntary, and are not perceived as coming from credible sources. Those interested in improving teaching recommend adopting a more formative developmental feedback model that endeavors to improve performance on a task (Weimer and Lenze, 1994) and solicits volunteers who have been shown to be more receptive to receiving feedback (Blumenthal, 1978; Sweeney and Grasha, 1979). “Feedback seeking” is the better description for this type of situation, because individuals are motivated to voluntarily seek feedback for their own improvement (Ashford et al., 2003).

Organizational psychologists characterize two major competing motives that influence the likelihood that someone will voluntarily seek feedback related to job performance. Ashford and colleagues explain that “individuals are instrumentally motivated to obtain valued information, but are also motivated to protect and/or enhance their ego and to protect others’ impressions of them” (Ashford et al., 2003, p. 774). Perceived benefits and costs are weighed in each decision. For perceived benefits, feedback seekers look for credibility, seeking feedback from individuals who possess relevant and accurate information (Fedor et al., 1992; Finkelstein and Fishbach, 2012). Negative feedback is accepted only if it comes from a high-status source (Ilgen et al., 1979), and status changes both the perception of and the desire to respond to feedback (Ilgen et al., 1979; Greller, 1980). On the other side, costs to one’s ego are also considered. For example, researchers find that individuals with longer time on the job seek less feedback, possibly due to reduction in perceived value or increased perception of costs (Ashford, 1986). In addition, feedback is more likely to be sought if the situation is uncertain and the individual perceives the risk to his or her job warrants this sacrifice of his or her ego (Hays and Williams, 2011). Individuals are more likely to seek feedback if the supervisor shows respect and concern (VandeWalle et al., 2000) and if the feedback will be private and developmental rather than public and evaluative (Ashford and Northcraft, 1992).

The organizational context for university faculty bears some similarity to the corporate and K–12 scenarios studied above. Our tiered system of ranks denotes status, and established individuals with tenure have less uncertainty about
their future than junior faculty and instructors. One of the major differences may be the particularly low value associated with job performance in teaching and the associated lack of reward for these activities (Hativa, 1995; Walczyk and Ramsey, 2003; Gibbs and Coffey, 2004; AAAS, 2010; Mervis, 2013). Faculty members attribute greater value to feedback if it comes from sources who are knowledgeable, and they also consider the perspective and motivation of the source (Wergin et al., 1976). Applying the principles from an organizational setting, one would predict that junior university faculty would be more likely to voluntarily seek out feedback if it is perceived as providing value—for example, increasing likelihood of receiving tenure and promotion. Feedback would also be accepted (even negative feedback) and responded to if the source is in a position of greater status. For tenured faculty members, there is less value added from feedback. They are not likely to gain status as a result of improving their teaching, so the cost to their self-image may be too great to warrant voluntarily seeking feedback from peers.

Vision for Feedback in Higher Ed

We summarize here these research findings to help formulate specific suggestions for structuring feedback (Table 1). That way, feedback may be structured to best support a faculty recipient in modifying and improving his or her teaching. If at all possible, feedback should be delivered immediately and on more than just one occasion. This could entail going over instructional materials before a class and immediately discussing thoughts for improvement, or right before and after a class session, but not after the long delay common to end-of-semester student evaluations or peer evaluations. Feedback providers need to be perceived as sympathetic, credible, and unbiased. Selecting coaches from outside the tenure-granting department may minimize conflicts and preserve collegiality and allow senior faculty access to expert role models (Huston and Weaver, 2008). However, research from peer coaching in the K–12 setting using collaborative teams of teachers of equal status rather than expert supervisors also showed demonstrable improvement on changing teaching behavior and student achievement (Showers, 1984). Stes and colleagues (2010), reviewing the handful of studies empirically examining the effects of instructional mentoring or coaching in higher education, noted an increase in teacher attitudes (Finkelstein, 1995; Gallos et al., 2005; McShannon and Hynes, 2005) and knowledge (Harnish and Wild, 1993) after peer mentoring and coaching. However, none of these studies utilized comparison groups or empirically and specifically tested the effect of the mentor’s status. Regardless of their status, providers need to be able to account for individual differences in experience and self-confidence when counseling recipients. To be most useful, feedback should be voluntarily sought. Newer faculty members, be they tenure-track or not, are more likely to appreciate the benefit of feedback to their advancement. Senior faculty members without the need to achieve promotion may respond better to encouragement and attaining goals such as documenting improved student learning in their classes. Finally, the most effective feedback identifies errors in a positive manner and provides examples of how to improve. This requires an increased openness and visibility where it is accepted that faculty regularly observe teaching in the classroom in the same manner used when mastering a new research technique. It also requires better descriptions (task standards) that explain what evidence-based practices look like during implementation (i.e., the taxonomy of observable practices for scientific teaching in development by Swarts et al., 2013).

OVERCOMING EXISTING BARRIERS: STRATEGIES FOR RECIPIENTS OF FEEDBACK

In this section, we identify barriers to implementing best practices for providing effective feedback on undergraduate teaching. Then we highlight strategies that recipients of feedback may borrow from existing programs facilitating pedagogical change and faculty development.

Situational barriers to providing effective feedback are apparent early in faculty careers. In fact, these barriers begin in graduate school. During their graduate training, most faculty members had few opportunities for teacher development: only a third of science graduate students report having access to a one-semester training in pedagogy (Golde and Dore, 2001; Tanner and Allen, 2006). Given this lack of professional development, many instructors are unaware of pedagogical techniques (Crouch and Mazur, 2001; Handelsman et al., 2004; Pukkila, 2004; DeHaan, 2005). Therefore, it is unsurprising that effective use of challenging pedagogical techniques is rare (Andrews et al., 2011; Henderson et al., 2012). This lack of training ultimately impacts not only the use of good teaching practices but also ability to provide instructional feedback. Scientists' professional identities may also act as a barrier to widespread reform in science education, an idea proposed by Brownell and Tanner (2012). Teaching is sometimes an undervalued part of faculty professional identity: Incorporating long-term ongoing opportunities for pedagogical development for graduate students can address this barrier by promoting innovative ways to seek and give feedback at the earliest stages of faculty careers (Brownell and Tanner, 2012).

Alternatively, faculty members may be aware of evidence-based teaching methods, but demonstrate a performance gap between what they are doing (or not doing) as compared with what they should be doing (Andrews et al., 2011; Ebert-May et al., 2011). After exposure to these teaching practices at workshops, faculty may need additional support through implementation (Table 2). While discipline-based science education research continues to grow, there are not necessarily in-house experts to provide feedback in each department, and these individuals may not have sufficient status for their feedback to be valued. Showers’ model (1984) supports the hypothesis that peers can be effectively trained as coaches, and Bernstein (2008) mentions several models for engaging centers for teaching and learning and fellow faculty members in the process (Hutchings, 1995; Chism, 2007). Buy-in to evidence-based teaching practices may be another barrier, however. Faculty may be resistant to change, for reasons such as commitment to content coverage (Anderson, 2002), lack of confidence in student ability (Brown et al., 2006; Henderson and Dancy, 2007), employment as adjunct faculty with different expectations and campus involvement (Roney and Ulerick, 2013), or concerns over classroom management (Welch et al., 1981). Consequently, instructional feedback may not be framed from a reformed perspective.
Moreover, the reward structure at research institutions often undervalues teaching (Hativa, 1995; Walczyk and Ramsey, 2003; Gibbs and Coffey, 2004; AAAS, 2010; Mervis, 2013). Often, there are no formal mechanisms in place for offering peer feedback beyond promotion and tenure evaluations, nor rewards for participating in a peer-feedback process. Faculty members may lack incentives for improving teaching while facing high expectations for research productivity (Boyer Commission on Educating Undergraduates in the Research University, 1998; NRC, 2003; DeHaan, 2005). Taken together, these barriers compound over time so that a sense of community around teaching in higher education may not be the norm.

Given the barriers described above, we recommend that change-makers and faculty development consultants consider the following example. In nursing, researchers have identified a systematic approach to improve productivity and competence (Stolovitch et al., 2000). The stepwise approach involves first analyzing the performance gap to understand the difference between the behavior exhibited and expectations, as well as its significance. Then, the underlying cause of the gap is identified before an appropriate intervention is selected. Finally, subsequent change is measured (Stolovitch et al., 2000). This approach has relevance for higher education, as there may be multiple underlying reasons that faculty may fail to adopt evidence-based teaching practices. Feedback providers should use knowledge of the reason(s) why someone is not implementing evidence-based teaching practices to frame and develop appropriate feedback interventions. Change-makers should consider that multiple intertwined causes may prevent effective implementation. This stepwise analysis supports feedback-giving efforts tailored to individuals’ needs and challenges with room for flexibility, variation, and change.

Both change-makers and feedback recipients might look to strategies that support shifts in professional identity while building a sense of community around teaching (thus changing culture) (Table 2). Establishing faculty learning communities for those willing to participate may be one avenue for offering and receiving regular feedback to support faculty with feedback beyond student evaluations and drop-in peer evaluations. Peer coaching is another strategy that may support this shift. A peer-feedback model, unlike a one-time classroom observation, is all-encompassing—providing feedback about everything from learning objectives to assessment strategies—rather than just evaluating the in-class performance. In this model, instructors regularly observe one another, providing support, feedback, and assistance in order to improve one another’s instructional practices (Mallette et al., 1999; Weimer, 2002; Huston and Weaver, 2008). Weimer (2002, p. 197) suggests that this is a way to let peers “function as colleagues and work collaboratively on improvement efforts.” Weimer offers two recommendations that are useful guiding principles: first, practice the “golden rule” in giving feedback, “giving feedback, ‘give unto each other the kind and quality of feedback you would like to receive,’” and second, develop an agenda. With a defined agenda, faculty members may learn and reflect together on specific problems. This shifts the feedback giving-and-receiving dynamic from a one-way exchange to more productive two-way communication. Both faculty learning communities and peer coaching may support science, technology, engineering, and mathematics (STEM) faculty grappling with student resistance to evidence-based instructional practices.

| Table 2. Resources for providing feedback in higher education |

<table>
<thead>
<tr>
<th>Type</th>
<th>Resources for feedback in higher education</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conferences and workshops</td>
<td>- Instructional development workshops (centers for teaching and learning, National Academies Summer Institutes [<a href="http://www.academiessummerinstitute.org">www.academiessummerinstitute.org</a>])</td>
</tr>
<tr>
<td></td>
<td>- Process Oriented Guided Inquiry Learning (<a href="https://pogil.org">https://pogil.org</a>)</td>
</tr>
<tr>
<td></td>
<td>- Project Kaleidoscope meetings (PKAL, American Association of Colleges and Universities [<a href="http://www.auco.org/pkal">www.auco.org/pkal</a>])</td>
</tr>
<tr>
<td></td>
<td>- CIRTL (<a href="http://www.cirtl.net">www.cirtl.net</a>)</td>
</tr>
<tr>
<td>Online videos</td>
<td>- iBiology education videos from the American Society for Cell Biology (<a href="http://www.ibiology.org/ibioeducation.html">www.ibiology.org/ibioeducation.html</a>)</td>
</tr>
<tr>
<td></td>
<td>- Howard Hughes Medical Institute biological demonstrations (<a href="http://www.researchandteaching.bio.uci.edu/lecture_demo.html#ATP">www.researchandteaching.bio.uci.edu/lecture_demo.html#ATP</a>)</td>
</tr>
<tr>
<td>Classroom observation protocols</td>
<td>- Classroom observation protocols (RTOP [<a href="http://physicised.buffalostate.edu/AZTEC/RTOP/RTOP_full/about_RTOP.html">http://physicised.buffalostate.edu/AZTEC/RTOP/RTOP_full/about_RTOP.html</a>])</td>
</tr>
<tr>
<td></td>
<td>- Classroom Observation Protocol for Undergraduate STEM (COPUS; Smith et al., 2013)</td>
</tr>
<tr>
<td></td>
<td>- Taxonomy of observable practices for scientific teaching (Swarts et al., 2013)</td>
</tr>
<tr>
<td></td>
<td>- Electronic Quality of Inquiry Protocol (EQUIP; Marshall et al., 2010)</td>
</tr>
<tr>
<td>Departmental culture</td>
<td>- Discuss expectations of department and methods for dealing with student resistance (Seidel and Tanner, 2013)</td>
</tr>
<tr>
<td></td>
<td>- PULSE Vision &amp; Change Rubrics (Aguirre et al., 2013)</td>
</tr>
<tr>
<td>Peer evaluation</td>
<td>- Excellent peer evaluation of teaching guide at <a href="http://tennltc.utk.edu/ut-peer-evaluation-of-teaching-guide">http://tennltc.utk.edu/ut-peer-evaluation-of-teaching-guide</a></td>
</tr>
<tr>
<td></td>
<td>- Peer review of teaching project <a href="http://www.courseportfolio.org/peer/pages/index.jsp">www.courseportfolio.org/peer/pages/index.jsp</a></td>
</tr>
<tr>
<td></td>
<td>- Peer Review of Teaching: A Sourcebook, 2nd ed. (Chism, 2007)</td>
</tr>
<tr>
<td></td>
<td>- &quot;The role of colleagues in the evaluation of college teaching&quot; (Cohen and McKeachie, 1980)</td>
</tr>
</tbody>
</table>

Given what we know about best practices for feedback, we recommend that change-makers, feedback providers, and feedback recipients focus on identifying how to make feedback specific, timely, corrective, and positively framed. Both change-makers and feedback recipients might borrow tools from existing faculty development programs to structure higher-quality feedback (Table 2). For example, interested faculty might use the feedback practices used by the Faculty Institutes for Reforming Science Teaching (FIRST IV; www.msu.edu/~first4/index.html). FIRST IV participants watch videotaped classroom sessions and then respond to questions such as: "What are the students doing? What
is the instructor doing? How would you go about changing this classroom so it is more student-centered? What is the instructor doing that students themselves should be doing?” Participants discuss and reflect, and then perform self-evaluations of their own videotaped classroom samples in concert with peer and expert review. Faculty may use rubrics developed by the Partnership for Undergraduate Life Science Education (PULSE; www.pulsecommunity.org). These rubrics are intended to structure departmental-level discussion and reflection about how program curricula and teaching practices align with Vision and Change goals. Faculty may use these rubrics to spark more nuanced discussions about feedback for teaching practices. Extensive additional resources are available through the Center the Integration of Research, Teaching, and Learning (CIRTL; www.cirtl.net) and the Measures of Effective Teaching (MET) project (www.metproject.org/faq.php).

Feedback recipients may be their own best advocates for receiving more useful feedback (Tables 1 and 2). Feedback recipients could propose a preclassroom observation meeting to discuss class goals, challenges faced, and areas for a peer observer to suggest specific strategies. This preobservation meeting may set up a framework for feedback recipients to receive more thoughtful, focused, practical feedback. Such a framework may also increase feedback recipients’ perception of the value of feedback and give them a voice in the process. Because barriers to accessing locally based learning communities may exist, programs such as PULSE make use of technology to share resources across institutional borders. We encourage feedback recipients to think beyond their department walls, to seek additional feedback from external mentors. From research about highly effective athletic coaches, we know that individuals with strong social networks who discussed their practices with others and dedicated portions of their off-season to studying their sports, had larger winning records than coaches who did not (Horton and Young, 2010). Essentially, winning coaches were more successful because they actively sought out feedback to improve their performance. Instructors, like coaches, also benefit from discussing their practices and sharing feedback to achieve a winning season as measured by student achievement. This mirrors what we know about how people learn: we continually reconstruct our understanding of the world and this process is social (Bransford et al., 2000). Likewise, we need to actively seek feedback to revise and improve our teaching practices.

AREAS FOR FURTHER RESEARCH

What we know about best practices for feedback primarily comes from the realm of K–12 teacher education research, as well as organizational psychology research. Research about best practices for instructional feedback in higher education—for college faculty—is uncharted territory. Here, we propose several areas of instructional feedback in need of more research, specifically focusing on instructional feedback for college faculty and potential outcomes related to student experiences.

Many faculty members, including educational researchers, are confused or disagree as to what exactly constitutes active learning (Hativa, 1995; Miller et al., 2000; Winter et al., 2001; Hanson and Moser, 2003; Yarnall et al., 2007; Chi, 2009; Allendoerfer et al., 2012). As a result, faculty members struggle to define the standards by which to frame feedback. Few models exist, consequently even faculty members who have attended workshops about active learning mischaracterize their own performance (Ebert-May et al., 2011). This disconnect between understanding and implementation suggests that feedback must clarify specific expectations while limiting contradictory information. One resource compilation to help instructors better envision and create engaged classroom environments is in development: the iBiology Project at the American Society for Cell Biology is in the process of creating and posting videos through their iBiology Education YouTube channel that showcase evidence-based classroom practices (iBioEducation, 2013). Research is needed to address questions such as: How does feedback that includes clarification about effectively implementing evidence-based teaching practices impact faculty teaching practices? In other words, to what extent does “clarifying the task” aid instructors? Does this increase the likelihood that faculty members are able to accurately define and effectively implement active-learning strategies?

We know that simply providing instructors with evidence about their teaching practices is not enough to instigate improved teaching (Andrews and Lemons, personal communication). Tools are needed to provide structured feedback for evidence-based teaching practices that will both support implementation and inform a peer-teaching evaluation system. Classroom observation protocols exist (e.g., the Reformed Teaching Observation Protocol; Sawada et al., 2002), but these are used for evaluative research purposes rather than for formative feedback, and the measurement scales are challenging to interpret (Marshall et al., 2010). Moreover, these do not offer strategic feedback for improvement (Marshall et al., 2011). New classroom observation protocols are in development that may be useful for formative instructional feedback (Eddy et al., 2013; Smith et al., 2013; Swarts et al., 2013), as is a feedback tool to improve evidence-based teaching practices (Gormally et al., unpublished data). More work is needed to understand: What are effective means for instructional feedback in higher education? How should this feedback be structured? What types of feedback do instructors report as most engaging them in trying new techniques?

To understand how to motivate faculty to seek and use feedback, we need to clarify the types of feedback desired by faculty in different job settings. First, we need to know more about how faculty members give and receive feedback. Then we can question whether informal or formal feedback approaches yield different outcomes in terms of how instructors perceive and respond to the feedback. How does an instructor’s perception of a feedback provider’s value impact his or her response to feedback? How does the manner in which the feedback is conveyed impact instructor morale? How do different types of faculty respond to different ways of conveying instructional feedback? It will also be critical to characterize, measure, and quantify instructional change following feedback. How do faculty behaviors, beliefs, and attitudes change as a result of feedback? How do faculty professional identities shift as a result of feedback? Researchers may explore whether we begin to see a cultural shift and whether “what gets rewarded, gets done” will encompass both research and teaching.
Studies show modest but significant improvements in teaching as measured by student perceptions (through student evaluations) of faculty change (Cohen, 1980; Safavi et al., 2013). We need to understand whether receiving feedback ultimately impacts student outcomes. How do students perceive changes in teaching behaviors following feedback? Further, how might end-of-semester course evaluations be revised to be more learner centered? How might the type of feedback elicited from a learner-centered course evaluation differ from a teacher-centered course evaluation? Do faculty members view this feedback as more valuable than traditional teacher-centered course evaluations? Do more faculty members report using this feedback? How might this feedback be used to address or head off student resistance in future courses? How does feedback lead to change that impacts student attitudes about the classroom environment, pedagogy, and learning science? Research to address these questions could substantially affect both faculty and student resistance to adopting evidence-based practices.

People are more likely to increase effort when “the goal is clear, when high commitment is secured for it, and when belief in eventual success is high” (Kluger and DeNisi, 1996). The efforts on the part of STEM instructors to reform instruction and shift the status quo closer to evidence-based teaching practices are heroic and ongoing, but we must match these efforts with improved instructional feedback. More research is needed to understand the outcomes and impacts of offering feedback to faculty. Implementing a reformed instructional feedback protocol, in addition to reformed teaching, may seem daunting. However, our current strategies for providing instructional feedback in STEM are inadequate. Therefore, we must challenge one another to move beyond student evaluations and the typically unproductive drop-in observations. Instead, we must advocate for more research in STEM education that focuses on the outcomes of improved instructional feedback, leading to the development and implementation of successful models of instructional feedback.

ACKNOWLEDGMENTS

The authors acknowledge continuing support and feedback from the University of Georgia Biology Education Research Group. This work was supported by National Science Foundation grant DUE-0942261 to PB.

REFERENCES


Centra JA. (1993). Reflective Faculty Evaluation: Enhancing Teaching 
and Determining Faculty Effectiveness, San Francisco, CA: Jossey-
Bass.

Centra JA. (2000). Evaluating the teaching portfolio: a role for col-
leagues. New Direct Teach Learn 2000(83), 87–93.

Chandler TA. (1978). The questionable status of student evaluations 
of teaching. Teach Psychol 5, 150–152.

Cheng DA. (2011). Effects of class size on alternative educational 


Chi MTH. (2009). Active-constructive-interactive: a conceptual fram-
ework for differentiating learning activities. Top Cogn Sci 1, 73–105.

Bolton, MA: Anker.

Cohen PA. (1980). Effectiveness of student-rating feedback for im-
proving college instruction: a meta-analysis of findings. Res High 

Cohen PA, McKeachie WJ. (1980). The role of colleagues in the evalu-
ation of college teaching. Improving Coll Univ Teach 28, 147–154.

Cossairt A, Hall RV, Hopkins BL. (1973). Experiments of experimen-
ters instructions, feedback, and praise on teacher praise and student 

Coulter GA, Grossen B. (1997). The effectiveness of in-class instructive 
feedback versus after-class instructive feedback for teachers learning 


Dancy M, Henderson C. (2010). Pedagogical practices and instruc-

d’Apollonia S, Abrami PC. (1997). Navigating student ratings of in-

DeHaan RL. (2005). The impending revolution in undergraduate sci-

Derting TL, Ebert-May D. (2010). Learner-centered inquiry in un-
dergraduate biology: positive relationships with long-term student 
achievement. CBE Life Sci Educ 9, 462–472.

Dowell DA, Neal JA. (1982). A selective review of the validity of 


Ebert-May D, Derting TL, Hodder J, Momsen JL, Long TM, Jardeleza 
SE. (2011). What we say is not what we do: effective evaluation of 
faculty professional development programs. BioScience 61, 550–
558.

Eddy S, Converse M, Abshire E, Longton C. (2013). Development and 
implementation of an instrument to characterize active learning in 
large lecture classes. In: Conference Proceeding, ed. AP Wenderoth, 
Minneapolis, MN: Society for the Advancement of Biology Education Research.

Englert CS, Sugai G. (1983). Teacher training: improving trainee per-
formance through peer observation and observation system technol-
ogy. Teach Educ Special Educ 6, 7–17.

Fedor DB, Buckley MR. (1987). Providing feedback to organizational 

expected to affect feedback seeking—a longitudinal-field study. Pers 
Psychol 45, 779–805.

Feldman KA. (1988). Effective college teaching from the students’ and 
faculty’s view: matched or mismatched priorities? Res High Educ 28, 
291–344.

Outcomes of the Katz/Henry Faculty Development Model, South Orange: New Jersey Institute for Collegiate Teaching and Learning.

Finkelstein SR, Fishbach A. (2012). Tell me what I did wrong: ex-
erts seek and respond to negative feedback. J Consumer Res 39, 22–
38.

ratings: a closer look. Paper presented at the annual meeting of the 

Freeman S, O’Connor E, Parks JW, Cunningham M, Hurley D, Haak 
increases performance in introductory biology. CBE Life Sci Educ 6, 
132–139.

Gallos MR, van den Berg E, Treagust DF. (2005). The effect of inte-
grated course and faculty development: experiences of a university 
Educ 27, 985–1006.

Gibbs G, Coffey M. (2004). The impact of training of university teach-
ers on their teaching skills, their approach to teaching and the ap-
proach to learning of their students. Active Learn Higher Educ 5, 
87–100.

Giebelhaus CR. (1994). The mechanical third ear device: a student 

Golde CM, Dore TM. (2001). At Cross Purposes: What the Experi-
ences of Today’s Doctoral Students Reveal about Doctoral Education, 

Goldman L. (1993). On the erosion of education and the eroding foun-
dations of teacher education (or why we should not take student evalua-
tion of faculty seriously). Teacher Educ Q 20, 57–64.

Green G, Osborne JG. (1985). Does vicarious instigation provide sup-

Greller MM. (1980). Evaluation of feedback sources as a function of 


Hanson S, Moser S. (2003). Reflections on a discipline-wide project: 
developing active learning modules on the human dimensions of 

professional development strategy for faculty. Commum College J 

Hativa N. (1995). The department-wide approach to improving fac-
ulty instruction in higher-education—a qualitative evaluation. Res 
High Educ 36, 377–413.

77, 81–112.

Hays JC, Williams JR. (2011). Testing multiple motives in feedback 
seking: the interaction of instrumentality and self protection moti-

Henderson C. (2008). Promoting instructional change in new faculty: 
an evaluation of the physics and astronomy new faculty workshop. 

Henderson C, Beach A, Finkelstein N. (2011). Facilitating change in 
undergraduate STEM instructional practices: an analytic review of the 

Henderson C, Dancy MH. (2007). Barriers to the use of research-
based instructional strategies: the influence of both individual and 

Henderson C, Dancy MH. (2009). Impact of physics education re-
search on the teaching of introductory quantitative physics in the 


Is Peer Interaction Necessary for Optimal Active Learning?

Debra L. Linton,* Jan Keith Farmer,† and Ernie Peterson*

*Department of Biology, Central Michigan University, Mount Pleasant, MI 48858; †Department of Biology, Schoolcraft College, Livonia, MI 48152

Submitted October 17, 2013; Revised March 3, 2014; Accepted March 27, 2014

Meta-analyses of active-learning research consistently show that active-learning techniques result in greater student performance than traditional lecture-based courses. However, some individual studies show no effect of active-learning interventions. This may be due to inexperienced implementation of active learning. To minimize the effect of inexperience, we should try to provide more explicit implementation recommendations based on research into the key components of effective active learning. We investigated the optimal implementation of active-learning exercises within a “lecture” course. Two sections of nonmajors biology were taught by the same instructor, in the same semester, using the same instructional materials and assessments. Students in one section completed in-class active-learning exercises in cooperative groups, while students in the other section completed the same activities individually. Performance on low-level, multiple-choice assessments was not significantly different between sections. However, students who worked in cooperative groups on the in-class activities significantly outperformed students who completed the activities individually on the higher-level, extended-response questions. Our results provide additional evidence that group processing of activities should be the recommended mode of implementation for in-class active-learning exercises.

INTRODUCTION

Cooperative group-based active learning has become one of the benchmark techniques of the reformed, student-centered, introductory biology classroom. In these types of courses, students work together in groups during class to answer questions and solve problems related to key biological concepts. Many studies have shown the effectiveness of this technique in improving student learning (e.g., Udovic et al., 2002; Knight and Wood, 2005; Armstrong et al., 2007; Carmichael, 2009). Often, these types of studies of active learning compare student performance in cooperative group courses with performance in traditional lecture courses that include no student-centered activities. These types of studies have been crucial to the widespread acceptance of the value of active-learning techniques. However, this type of design does not allow us to identify the key components of an effective active-learning exercise. This is an important line of investigation that requires our attention. For active learning to become more widely practiced, as recommended by Vision and Change (American Association for the Advancement of Science, 2011), we need to make it more accessible to all instructors. Andrews et al. (2011) found that active learning, as practiced by randomly selected “typical” college biology instructors, was not associated with greater student learning compared with traditional lecture. They point out that most studies of active learning are predominantly conducted in the classrooms of science education researchers who have had extensive experience and training in these techniques. Andrews and colleagues suggest that a “rich and nuanced understanding of teaching and learning” must be developed by “typical” instructors in order for them to be similarly successful with implementing active learning. We can help provide this understanding by identifying the key elements in effective active-learning implementation.

An active-learning exercise typically begins with the instructor presenting a task to the class. Students then work
on the task. After the students have finished their work, some sort of full-class processing typically takes place. Implementation of each of these stages could vary in important ways. Many of these variations are highlighted in the published meta-analyses of active learning in science (e.g., Prince, 2004; Michael, 2006; Ruiz-Primo et al., 2011), and we have identified others based on our own experiences and conversations with colleagues. The nature of the question or problem itself varies widely in complexity. Students might be asked to answer a multiple-choice question (which themselves can vary in complexity and cognitive level), analyze graphical data, draw models, work through case studies, or perform a variety of other tasks. Students could work alone on the problem or in formal or informal cooperative groups. After group discussion, one team member might write the group’s answer or each team member might write his or her own answer or students may not be asked to write at all. Full-class processing may involve the use of clickers. The instructor may ask the students to report and share their ideas with the class for discussion or may simply explain the answer to the students (Turpen and Finkelstein, 2009). Given all the possible differences in implementation, it is no surprise the outcomes are not consistent. Even those who are experienced and successful with active learning do not all do it the same way. However, if we can identify which components are crucial to the successful implementation of active learning, we can make it less of a risk for new instructors (or “old” instructors who want to try something new) to give it a try.

We are interested specifically in the effect of peer interaction on student learning and self-efficacy. Social interdependence theory (Johnson and Johnson, 2009) leads us to predict that peer interaction and cooperation will increase student learning when positive interdependence exists. In addition to this enhancement of learning, Johnson and Johnson (2009) have determined that cooperative learning is positively related to a variety of psychological attributes, such as confidence, independence, and self-esteem. Self-efficacy (students’ belief in their ability to succeed) has been identified as an effective predictor of learning (Zimmerman, 2000). Therefore, we have included self-efficacy as a variable in this study.

Research in physics and chemistry education has found strong evidence that peer interaction can enhance learning. A meta-analysis (Gossler, 2011) of peer-led team learning (PLTL) in chemistry demonstrated increases in student learning in three different implementation formats. PLTL has also been shown to increase retention while maintaining rigor (Lewis, 2011). Peer instruction in physics has shown similar success (Crouch and Mazur, 2001) and is a widely accepted and practiced pedagogy in physics courses. However, Turpen and Finkelstein (2009) determined that differences in implementation of peer instruction can result in “disparate opportunities for students to practice conceptual reasoning, skills at talking physics, agency, and scientific inquiry” (p. 1).

Ruiz-Primo et al. (2011) identified differences in implementation of active learning and compared effect sizes in a meta-analysis of 197 comparative studies of innovations in science classrooms. In this analysis, they separated out different categories of innovations and their methods of implementation. The active-learning approach we have identified above would be categorized in their study as “conceptually oriented tasks + collaborative learning.” They found this combination to be the most common innovation infused into undergraduate science courses. They identified 41 published research studies implementing this approach (and collecting sufficient data for analysis) and found an effect size of 0.54. However, they also identified studies that implemented conceptually oriented tasks outside the structure of collaborative learning and found an effect size of 0.68. These results suggest greater gains when conceptually oriented tasks are implemented without cooperative learning.

We have to be careful when evaluating the effectiveness of peer interaction to make sure our results are not confounded by the effect of the activity that students experience. In the Ruiz-Primo et al. (2011) meta-analysis, studies that evaluated the use of conceptually oriented tasks versus traditional lecture are being compared with other studies that looked at cooperative activities versus traditional lecture. This is in contrast to a study design that could compare the use of the same conceptually oriented tasks implemented with and without collaboration. Studies of this type are needed for a variety of active-learning implementation issues.

Science education research is beginning to make progress on this agenda. Smith et al. (2009) found that peer discussion improved student performance on in-class clicker question sets. In 2011, Smith’s research team expanded this research to include differential implementation of the clicker activities. They found that peer discussion followed by instructor explanation lead to greater learning gains than either peer discussion or instructor explanation alone. If this finding is confirmed, then we can add this to our shared definition of effective active learning. Discrete investigations of this type, which isolate the effect of specific components, are needed to provide the evidence that builds a shared definition of effective active learning.

We were particularly interested in the idea that the implementation of conceptually oriented tasks might be more (or at least as) effective without the incorporation of peer interaction. It is possible that the learning gains that occur because of an activity may be due primarily to the “time on task,” individual thinking, or instructor explanation that occurs when practicing the application of recently learned biological concepts and may not require the cooperative group format. We do not believe that prior research has sufficiently disaggregated the effect of students’ interaction with the task from the effect of students’ interaction with their peers. We have attempted to do so here.

METHODS

Research Design

This research was approved by the Institutional Review Board (project #369307-1) at the university where it was conducted. Two sections of an introductory biology course for nonmajors were taught by the same instructor in the same semester, with back-to-back class meeting times. Both sections began the semester with 144 students enrolled; however, only 133 students in each section signed consent forms to participate in the research. The course included a 2-h laboratory experience each week. There were six laboratory sections of 24 students for each lecture section. The laboratory experiences, including all introductory materials and experimental
procedures, were the same for all sections. Consistency between sections was maintained by weekly meetings to train the graduate teaching assistants in facilitating each lab experience.

In both sections, 10- to 12-min lecture portions were alternated with in-class activities. In one section, students worked in cooperative groups on these activities. These groups were randomly assigned by the instructor, and each consisted of four students. The activities varied in complexity and included such tasks as: 1) drawing a model of a concept (e.g., relationship of nucleotide, gene, DNA, chromosome) or process (e.g., osmosis in freshwater vs. saltwater fish), 2) designing an experiment to test a hypothesis, 3) analyzing graphical data in context (e.g., changing phenotype frequency via natural selection), and 4) making a prediction or explaining a concept using evidence. In the other section, students completed the same activities but worked individually. All other aspects of course design, delivery, and assessment were held constant between the two sections. Full-class processing of each activity, facilitated by the instructor, occurred after the group or individual portion was completed. Varying balances of student and instructor explanation occurred during these discussions. However, in each case, the instructor ensured that the activity had been fully explained before moving on. The only difference between the two sections was the group versus individual processing of the activities.

For promoting positive group interdependence in the cooperative groups, one individual from each team was randomly selected to complete a five-question quiz at the end of each class period. All team members present in class that day were assigned the quiz grade this student earned. In the individual section, each student completed the quiz independently at the end of each class. The quiz score and points earned on the activity were compiled for an in-class score for each day.

At the beginning of the semester, students completed a demographic survey (e.g., gender, ethnicity, high school type, high school grade point average [GPA]), self-efficacy assessment instrument (Baldwin et al., 1999), and key concepts pretest to evaluate the equivalency of the two groups. The key concepts pretest included one question selected from the textbook (Campbell et al., 2006) test bank that the researchers and instructor agreed best represented the key concept of each of the 20 chapters covered during the semester, and modeled the format, quality, and cognitive level of questions that would be asked on the multiple-choice portion of the exams. The key concepts test and self-efficacy instrument were readministered at the end of the semester during the last regular class meeting. Student learning was assessed based on four exams, which consisted of both multiple-choice and extended-response questions. The texts of the essay questions that were analyzed for this research are included in Table 1.

### Table 1. Exam essay questions and Bloom ratings

<table>
<thead>
<tr>
<th>Exam</th>
<th>Bloom Level</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exam 1: Matter, Bloom level 3</td>
<td>Dry ice is the solid form of carbon dioxide. A completely sealed airtight container contains a piece of dry ice with a mass of 10 g. As the temperature rises, the solid dry ice is converted directly into gaseous CO₂. After the dry ice is completely converted to a gas, how has the amount of mass (matter) inside the container changed? Explain your answer using the vocabulary related to the nature of matter.</td>
<td></td>
</tr>
<tr>
<td>Exam 1: Protein production, Bloom level 2</td>
<td>Imagine a cell producing a protein that will be exported from the cell. Trace the pathway of production from the gene in the nucleus, to production of the protein, through the endomembrane system to the protein’s export from the cell. Name the organelles involved and what happens at each step.</td>
<td></td>
</tr>
<tr>
<td>Exam 2: Photosynthesis and cellular respiration, Bloom level 3</td>
<td>Explain how a carbon atom in a molecule of glucose in one of your muscles could become part of a glucose molecule in a tomato plant in your garden. As you trace the pathway, name both the processes involved and how the carbon atom exits and enters any organisms involved.</td>
<td></td>
</tr>
</tbody>
</table>
| Exam 3: Natural selection, Bloom level 4 | On the Hawaiian island of Kauai, there is a species of cricket named *Telogryllus oceanicus*. In the 1980s male crickets in this population could be heard chirping at night, attracting the females for mating. Almost all of the males chirped, but there was a small proportion of males that could not chirp because their wings lacked the structures needed for chirping. In the 1990s a fly species (*Ormia ochracea*) settled on the island. The flies implant their larvae in the bodies of crickets. The larvae use the crickets as food, and the crickets eventually die. The chirping males make so much noise that they are easy for the flies to locate, so they suffered the most mortality. By 2003, the proportion of nonchirping males had risen to more than 90%.

  a. How did the allele for nonchirping first come into existence?
  b. Explain how the process of natural selection resulted in the increase in the proportion of non-chirping males.
  c. If the flies were removed in 2004, predict what would happen to the proportion of chirping males in the cricket population over the next several generations. Explain your prediction. |
| Exam 4: Carbon cycling and energy flow in ecosystems, Bloom level 3 | This drawing [food chain of plant → grasshopper → mouse → snake → hawk] represents a food chain. Organic matter present in each organism represents both stored energy and stored carbon. The arrows represent movement of energy and carbon atoms through the food chain. Answer the following questions based on this drawing:

  a. Where did the energy stored in the organic matter originally come from and by what process did it enter the food chain?
  b. Explain the shortest route that a carbon atom in a glucose molecule in one of the mouse’s cells could take to end up back in a glucose molecule produced by the flower. For each step, name the processes involved and the form that the carbon will be in.
  c. What is the final fate of the energy stored in the mouse’s glucose molecule? |
Data Analyses
We ran chi-square analyses of each item on the student profile survey, comparing students in the cooperative group section with students in the individual activity section. Only students for whom we had key concepts pretests and all four sets of exam data were used in this analysis, as these are the only students we would be comparing in later analyses. The sample sizes for this analysis and all exam analyses are 102 students in the cooperative section and 98 in the individual section.

For an overall look at the effects on self-efficacy, scores on the self-efficacy instrument were analyzed by converting Likert responses to numerical values (totally confident = 5, through not at all confident = 1) and summing values to give a composite score for each student. Only students who completed both the presurvey and postsurvey were included in the self-efficacy analyses. This resulted in sample sizes of 92 students in the cooperative section and 95 students in the individual section. Pretest scores were compared between sections using a t test. Posttest scores were analyzed by analysis of covariance (ANCOVA), with scores on the pretest as the covariate.

Students’ percent scores on the key concepts pretest were compared between treatments using a t test. Posttest scores were compared using ANCOVA, with pretest scores as the covariate. Again, only students who completed both the pretest and posttest were included in the analyses. For this analysis, the sample sizes were 86 students in the cooperative section and 92 students in the individual section.

We rated each exam item by Bloom level, using the original Bloom’s taxonomy (1956). Three members of the research team rated each item. Any disagreements were discussed to reach consensus. Average Bloom levels were calculated for the multiple-choice and essay sections of each exam.

Students’ percent scores on only the multiple-choice portions of the exams were compared between sections using ANCOVA, with percent score on the key concepts pretest as the covariate.

Essay responses on the exams were assessed separately from the multiple-choice questions. Essay responses were coded for specific key correct concepts based on a coding rubric developed by the research team. The coding rubrics were developed by the lead investigator, then reviewed by the rest of the research team, and edited based on their feedback. The rubrics were then piloted on a random sample of student work and further modified. The final rubrics are shown within the data figures for the essay analyses below. Essay answers were deidentified before coding, so raters were not aware of a student’s treatment status while coding that student’s answers. Two raters coded each essay, and disagreements were settled by discussion. The total number of correct concepts included in student responses was compared between groups by ANCOVA, using the percent score on the key concepts pretest as the covariate. Individual concepts were analyzed for differences in frequency between sections using chi-square analyses.

RESULTS

Demographic Data
No significant demographic differences were detected between the two sections in gender, major, highest level of parents’ education, high school type (urban/suburban/rural or public/private/homeschool), previous biology experience, or biology GPA. However, a significant difference was found for overall high school GPA (p = 0.04), with students in the cooperative group section self-reporting slightly higher overall high school GPA (Figure 1). On the basis of this information, we decided to use scores on the key concepts pretest as a covariate in analyses to minimize any inherent differences between sections. The GPA data were not used as a covariate, because they were categorical (students selected from GPA ranges) and also because they were less reliable self-reported data.

Self-Efficacy
Scores on the self-efficacy pretest were not significantly different (p = 0.63) between sections. Scores on the posttest were also not significantly different (p = 0.35; Figure 2). However, both the cooperative (p = 0.00076) and individual (p = 0.016) sections showed significant increases in self-efficacy from pretest to posttest (Figure 2).

Figure 1. Self-reported high school GPA distributions of students in cooperative and individual sections. Distributions are significantly different (p = 0.04). n = 102 cooperative, 98 individual.

Figure 2. Average self-efficacy scores, pretest and posttest for both treatments. Bars represent SE. Means are significantly different pretest vs. posttest for both the cooperative (p = 0.00076) and individual (p = 0.015) sections, but not cooperative vs. individual on either the pretest (p = 0.63) or posttest (p = 0.33). n = 92 cooperative, 95 individual.
Fifty-nine separate chi-square analyses were run on individual items from the exam essay rubrics. Seventeen items were found to have significant differences between the cooperative and individual sections. In each of those 17 cases, the cooperative section showed a significantly higher percentage of students including that item in their essay answer than did the individual section.

On exam 1, two essay questions were analyzed focusing on the major concepts from the chemistry and cell unit. Students in the cooperative section were more likely to reference the law of conservation of matter (Figure 6) in their explanation of the nature of matter and were also more likely to explain the function of the Golgi apparatus in the protein production (and modification) pathway (Figure 7).

On exam 2 (Figure 8), the key essay question related to the movement of carbon atoms during photosynthesis and cellular respiration. Five items were significantly different between sections. Students in the cooperative group section were significantly more likely to explicitly state that carbon dioxide is released from the body and subsequently taken in by a plant. They were also significantly more likely to identify the processes of cellular respiration and photosynthesis as the driving forces for this movement of carbon in ecosystems.

On exam 3 (Figure 9), the major essay featured natural selection. There were six concepts that students in the cooperative section included significantly more frequently than did students in the individual section. These concepts included genetic variation, differential survival, differential reproduction, and correct use of the term "fitness" in two different contexts.

The essay question analyzed for exam 4 focused on the flow of energy and cycling of carbon in ecosystems. Four concepts showed significant differences between sections (Figure 10), with students in the cooperative section being more likely to describe the production of carbon dioxide by cellular respiration and to explicitly state that the carbon for photosynthesis comes from the atmosphere and enters a plant through stomata and that the eventual fate of energy in ecosystems is to be lost as heat.

**DISCUSSION**

In this study, peer interaction was not shown to have a significant influence on students’ overall self-efficacy. Self-efficacy increased from the beginning to the end of the semester, but students in the cooperative groups did not realize any greater overall increase than students who worked individually. This is in contrast to data from Prince’s (2004) meta-analysis, which found several studies showing improved “self-esteem” from collaborative and cooperative learning compared with traditional lecture. Citing the work of Johnson et al. (1998) and a meta-analysis by Springer et al. (1999), they found increases in students’ self-esteem and attitudes, with effect sizes of 0.29–0.67, in collaborative or cooperative groups. Similarly, Fencel and Scheel’s 2005 study showed that collaboration can make positive contributions to self-efficacy. However, recall that our “noncooperative” treatment was not traditional lecture and self-efficacy did significantly increase in both sections. Individual student engagement with conceptually oriented tasks showed as much of an impact on self-efficacy as did...
group processing of those same activities. This suggests that peer interaction is not the primary driver of the increases in self-efficacy that have been seen in active learning.

Our results did not show any effect of peer interaction on students’ ability to correctly answer multiple-choice questions, either on the key concepts posttest or on the exams. Previous studies have shown that peer discussion can improve student performance on multiple-choice questions. In the study conducted by Smith et al., (2009), students discussed multiple-choice questions in cooperative groups and showed improvement on later multiple-choice questions on the same concepts. However, in their design, the improvement was not compared with any other mode of interacting with the questions. Smith and colleagues’ research was investigating whether peer discussion can help students learn, while we are asking whether peer discussion is necessary to maximize learning during active-learning activities or whether students could interact in other ways and receive equal or greater benefit. Recall that in our study, full-class processing of the activities facilitated by the instructor always occurred after the group or individual work. This was not the case in Smith et al.’s 2009 paper. Later research by Smith and colleagues (2011) compared student performance on clicker questions under three scenarios: peer discussion alone, instructor explanation alone, and peer discussion followed by instructor explanation. This research is more analogous to our study. In our cooperative sections, group work was always followed by instructor-facilitated explanation. In the individual section, there was only instructor-facilitated explanation. Smith et al. (2011) found that peer discussion followed by instructor explanation led to greater increases in performance than did instructor explanation alone. On the multiple-choice part of our research, this was not the case.

The comparison of student performance on essay questions tells a different story. Students in the cooperative section included more correct concepts in their answers than did students in the individual section. When individual concepts were compared, many significant differences were found, and in all cases in which a significant difference was found, the cooperative group section had a higher proportion of students including these key concepts in their essay answers. The fact that there were no differences seen in multiple-choice performance makes this trend stronger, as it (along with the use of key concepts pretest scores as covariates in all analyses) weakens any argument that students in the cooperative section were simply better students. The evidence is strong that working in cooperative groups and engaging in peer discussion resulted in greater performance by students on the higher-level assessments. The multiple-choice questions on the exam required primarily low-level cognitive skills as rated using Bloom’s taxonomy. This is a common pattern in introductory biology courses (Momsen et al., 2010). However,
the essay questions did make higher cognitive demands on students. The in-class activities used in the course varied in cognitive level but included many activities that would be rated at Bloom’s levels 3–5. We suggest that the benefits that arise from peer interaction are most likely seen at these cognitive levels. Students do not need their peers to help them memorize a fact and repeat it back on the exam. This finding aligns with Johnson and Johnson’s (2009) predictions based on social interdependence theory. They identified conditions under which individual efforts might be most appropriate. One of the conditions they identified was “Unitary, nondivisible, simple tasks need to be completed, such as the learning
Figure 8. Coding rubric and percent of students in each section including each concept for the photosynthesis/respiration essay question on exam 2. The text of the question is shown in Table 1. Significant differences ($p < 0.05$) between sections are indicated by $^*$. 

Figure 9. Coding rubric and percent of students in each section including each concept for the natural selection essay question on exam 3. The text of the question is shown in Table 1. Significant differences ($p < 0.05$) between sections are indicated by $^*$. 

D. L. Linton et al.
of specific facts or the acquisition or the performance of simple skills” (p. 370). Our results provide evidence in support of this hypothesis.

However, cognitive tasks that require application, analysis, and synthesis benefit from the experience students have discussing these concepts with their peers. On the basis of the example clicker questions included in Smith et al.’s 2011 paper, we do not see their data as contradictory to ours. Their multiple-choice questions were not the simple knowledge and understanding (Bloom’s levels 1–2) questions typical of introductory biology exams, but required higher-level processing skills. From this perspective, the two studies support each other. Similarly, the ConcepTests and analogous exam questions used in physics peer instruction are not questions that can be answered by memorization (Crouch and Mazur, 2001). The proven effectiveness of peer instruction in these conditions further supports our findings. On higher-level tasks, group discussion followed by instructor-facilitated explanation leads to optimal student learning.

Another trend that has captured our interest relates to the types of concepts that show significant differences between sections on the essay rubrics. For many of the items on the rubrics, there were key terms that raters were instructed to look for, for example, “conservation of matter,” “cellular respiration,” “photosynthesis,” “stomata,” and “fitness.” Other items focused on concepts that did not have specific terms associated with them. Many of the items that targeted key words showed significantly higher inclusion by students in the cooperative group section. The current study was not designed to test this idea, so we cannot say anything definitive about this but suggest that it could be an area for further study. Cooperative group discussions may encourage students to use correct vocabulary, as it improves communication skills by encouraging scientific discussion between peers.

In conclusion, the evidence we have presented supports the assertion that “peer interaction is necessary for optimal active learning.” However, the beneficial effects of peer interaction may only be seen on higher-order activities/assessments. The more complete answers written by students in the cooperative section suggest that it is not just the activity itself that improves learning but that peer interactions play a crucial role in promoting students’ understanding or at least students’ ability to explain their understanding. Therefore, as we begin to develop a shared definition of effective active learning, we propose that peer interaction, followed by instructor-facilitated explanation (during full-class processing), become a key component of that definition. While peer interaction followed by instructor-facilitated explanation is already a typical component of how most active learning is implemented, the framework of “scientific teaching” (Handelsman et al., 2004) requires us to make these decisions based on evidence. This study, coupled with Smith et al. (2011), builds on the foundational research on cooperative learning to provide strong evidence in favor of this approach within a conceptually oriented task framework utilizing higher-level
activities and assessments. We renew our call for more research of this type to identify other key components of effective active learning, so we can more easily and broadly disseminate this approach in a manner that will allow all instructors to become effective with it.

REFERENCES

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


Article

Integrating Quantitative Thinking into an Introductory Biology Course Improves Students’ Mathematical Reasoning in Biological Contexts

Susan Hester,* Sanlyn Buxner,† Lisa Elfring,* and Lisa Nagy*

*Department of Molecular and Cellular Biology and †Department of Teaching, Learning and Sociocultural Studies, University of Arizona, Tucson, AZ 85721

Submitted July 19, 2013; Revised October 2, 2013; Accepted October 6, 2013
Monitoring Editor: John Jungck

Recent calls for improving undergraduate biology education have emphasized the importance of students learning to apply quantitative skills to biological problems. Motivated by students’ apparent inability to transfer their existing quantitative skills to biological contexts, we designed and taught an introductory molecular and cell biology course in which we integrated application of prerequisite mathematical skills with biology content and reasoning throughout all aspects of the course. In this paper, we describe the principles of our course design and present illustrative examples of course materials integrating mathematics and biology. We also designed an outcome assessment made up of items testing students’ understanding of biology concepts and their ability to apply mathematical skills in biological contexts and administered it as a pre/postcourse test to students in the experimental section and other sections of the same course. Precourse results confirmed students’ inability to spontaneously transfer their prerequisite mathematics skills to biological problems. Pre/postcourse outcome assessment comparisons showed that, compared with students in other sections, students in the experimental section made greater gains on integrated math/biology items. They also made comparable gains on biology items, indicating that integrating quantitative skills into an introductory biology course does not have a deleterious effect on students’ biology learning.

INTRODUCTION

Over the past several years, there has been a call to revolutionize undergraduate biology education to reflect the field of modern biology. A recurrent theme throughout the resulting recommendations and guidelines is that undergraduate biology students must develop quantitative skills and learn to apply them in biological contexts (National Research Council [NRC], 2003, 2009; Association of American Medical Colleges and Howard Hughes Medical Institute [AAMC/HHMI], 2009; Labov et al., 2010; Woodin et al., 2010). Among the conclusions reached in the National Research Council’s BIO2010 and A New Biology for the 21st Century reports was that greater integration of quantitative skills into introductory biology classes is vital to the ongoing success of graduates from biology programs, as well as to recruitment of quantitatively talented students into biological fields (NRC, 2003, 2009). At the same time, the medical community has been revising its list of qualifications for future physicians. The AAMC/HHMI 2009 report, Scientific Foundations for Future Physicians, identified the ability to “apply quantitative knowledge and reasoning” as a key competency for those entering medicine (AAMC/HHMI, 2009).

Traditionally, life sciences students rarely experience mathematics within the context of their own disciplines.
Indeed, anecdotal evidence suggests that students who perceive themselves as math-weak gravitate toward biology, because they consider biology to be relatively math-free. These attitudes persist, despite mathematics course requirements for most life sciences students, because failure to integrate mathematics in meaningful ways for life sciences majors contributes to the perception that mathematics is irrelevant outside its discipline (Zan et al., 2006). Thus, in the minds of many biology students, mathematics and biology remain in two distinct, separate compartments. Increasing the quantitative thinking that students encounter in their biology courses may be a more effective way to challenge this perception and encourage transfer of mathematics skills than increasing traditional mathematics requirements for biology students alone.

Among its recommendations, the BIO2010 report encouraged exposing life sciences majors to quantitative thinking in their first year of study. Quantitative problem solving is a skill, and a difficult one to learn and master—if we want students to use quantitative problem solving in their advanced classes, we should lay the foundation for their success in introductory classes. The stakes are high when students first encounter math in biology in high-level classes in which both the biology concepts and mathematics skills are advanced. Introducing quantitative thinking earlier in the biology curriculum may increase the quantitative competency of students entering more advanced courses, paving the way for greater overall development of students' quantitative abilities by the time they complete their undergraduate course work. It may also give life sciences students a chance to reconsider their planned undergraduate training in math, an opportunity that will be lost to many students if their first exposure to math in a biological context occurs during their junior or senior year. Moreover, using quantitative thinking methods in introductory biology has the added benefit of exposing those students who will not continue in advanced biology courses to a better picture of modern biology as an integrated, analytical science (NRC, 2009). Few efforts to bring mathematics into introductory biology classes have focused on the development of online or written modules targeting particular skills (e.g., Robeva et al., 2010; Thompson et al., 2010). Our own previous efforts to develop and implement similar modules had disappointing results; we found that stand-alone modules, done on the students' own time, perpetuated the perception that math was an "add-on" that was not representative of the core content in biology (K. Dixon, personal communication). In light of this, we chose to follow a more integrated approach in which students would encounter brief in-class exercises in quantitative application throughout the semester that would repeatedly call on the same set of quantitative skills in different biological contexts. This approach is similar to that taken by Matthews and colleagues (2010) in developing an introductory survey course for biology majors, and by Speth and colleagues (2010) in developing an introductory course in genetics, evolution, and ecology. One way in which our approach differed, however, is that we primarily emphasized skills that the students were expected to have acquired previously in fulfilling the mathematical prerequisites for the course. Rather than introduce new quantitative skills, our goal was to encourage greater mastery and transfer of existing skills to biological contexts.

We designed an introductory molecular and cell biology course in which we targeted quantitative skills that we identified as supporting the biological concepts taught in the course and as being generally useful to students as they progress through their advanced biology courses. Throughout the course, we integrated quantitative and biological reasoning in course learning outcomes, in-class student activities, assignments, and exams.

In the process of initial course design, we defined a set of learning outcomes for introductory molecular and cell biology that encompassed both "traditional" concepts covered in such courses and the ability to apply a particular set of quantitative skills to biological problems. We designed a multiple-choice instrument intended to assess students according to a subset of those course learning outcomes. We administered the outcome assessment as a pre/postcourse test to measure the disparity between incoming students' grasp of biology concepts and their ability to apply quantitative skills along with those concepts, as well as to compare students in our experimental section of the course with their peers in other sections of the same course. Our precourse outcome assessment results confirm that students are unable to spontaneously transfer prerequisite mathematics skills to biological problems. Comparing pre/postcourse outcome assessment gains across sections showed students in the experimental section making greater gains in applying quantitative skills to biological problems than students in other sections, while making similar gains in biological concepts.

In the following, we describe the development of our outcome assessment instrument and our course, provide examples of course materials to illustrate our philosophy and methods, and present the results from administering our pre/postcourse outcome assessment.

MATERIALS AND METHODS

The courses described in this study were taught at the University of Arizona, a large state university (total student enrollment is ~39,000) located in Tucson. One of the authors (S.H.) instructed the experimental section of the course; another author (L.E.) instructed one of the large comparison sections using traditional (nonquantitative) learning outcomes. Study design and data collection were approved by an internal review board through the University of Arizona Human Subjects Protection office.

MCB181R: Introduction to Molecular and Cellular Biology

MCB 181R: Introduction to Molecular and Cellular Biology is one of a pair of introductory biology lecture courses offered at the University of Arizona. In terms of student population and course content, MCB 181R is representative of introductory biology courses at state universities throughout the United States (Gregory et al., 2011). The majority of the roughly 2000 students per year who take the course—life sciences majors and those preparing for entry into medical and medicine-related fields—are precisely the students identified as needing greater exposure to mathematics and application of mathematics in biological contexts. Despite this, the content of MCB 181R has traditionally been and remains almost entirely qualitative.
MCB 181R is taught by several instructors each year. Instructors choose a common textbook for the course and typically cover similar topics (e.g., macromolecules, central dogma, cell structures), but do not teach from a common set of materials. The associated lab section, MCB 181L, is also independent of the lecture sections; students may take the lab section before, concurrent with, or after the lecture section. Thus, conceptual emphasis, degree of detail covered in particular topics, ordering of and pacing through the topics, and pedagogical styles and strategies differ greatly across the different lecture sections and between the lab and lecture sections.

**We Identified Biological and Quantitative Learning Outcomes for the Course**

Identifying biological learning outcomes for the course was relatively straightforward. Using syllabi and learning outcomes from experienced MCB 181 instructors, we assembled a set of molecular and cell biology learning objectives. The biological learning outcomes that we identified emphasize such concepts as “structure affects function” and topics such as macromolecules, enzyme-catalyzed reactions, central dogma, and genetics, which are typical of such courses (Gregory et al., 2011).

Establishing quantitative learning outcomes for the course required more consideration. To choose an appropriate set of skills to integrate into course materials, we began by considering skills that 1) students should have developed by fulfilling their mathematical prerequisites for the course and 2) integrated naturally into molecular and cell biology materials. Our motivation for these two criteria was that we wished to build quantitative learning outcomes that were accessible to the students and that supported the biological learning outcomes instead of being distracting and contrived. The skills that we identified were algebra and manipulating units; scale, exponents, and logarithms; reading and creating graphs and tables; and counting and probability. Once we had identified these skills, we discussed them with other MCB 181 instructors, instructors of other classes for biology- and medical-track students, and mathematics faculty interested in the integration of mathematics and biology. In these conversations, we verified that these faculty members felt it was reasonable to expect students to apply the skills we had identified, but were often frustrated by their students’ inability to do so in the contexts of their biology courses.

**We Designed a Pre/Postcourse Outcome Assessment to Measure Student Gains in Course Learning Outcomes**

Having established the biological and quantitative learning outcomes for the course, we designed an instrument to measure student gains for certain outcomes. We developed a 24-item outcome assessment targeting learning outcomes in four central topics of molecular and cell biology (meiosis, genetics, and inheritance; nucleic acid structure and function; enzymes, energy, and the reactions of life; and gene regulation and central dogma) and application of three quantitative skill areas (algebra and units, counting and probability, and analyzing graphs) in the context of these topics.² We modeled our instrument design after the Introductory Molecular and Cell Biology Assessment (IMCA; Shi et al., 2010), and four of the items in our final outcome assessment are adapted from the IMCA. We found it necessary, however, to develop additional items to address target concepts and competencies not addressed by existing concept inventories. Our outcome assessment was developed over five semesters (Fall 2010 to Spring 2012) before its use in Fall 2012. Each semester, we sought and received feedback from molecular and cellular biology (MCB) faculty, MCB 181R and other introductory biology instructors, upper-division MCB undergraduates, and MCB graduate students by asking them to critically evaluate the items. We also administered versions of the outcome assessment to students in several sections of MCB 181R over four semesters (Spring 2011 to Spring 2012) before Fall 2012. Our outcome assessment instrument is included in the Supplemental Material.

Our outcome assessment consists of 13 items addressing biological concepts alone (referred to as “Bio” items throughout this text) and 11 items requiring quantitative application in the context of the same biological concepts (referred to as “BioMath” items throughout this text). In developing the Bio items, we focused in particular on learning outcomes shared by all participating sections of MCB 181R; when seeking feedback from instructors about these items, we asked that instructors discuss whether their postcourse students should be able to answer the items correctly given the items’ content and wording. Concepts in each biological topic were addressed by at least three nonquantitative items and three of the four were addressed by three quantitative items (the fourth was addressed by two quantitative items). Each quantitative skill was applied in at least three items. Table 1 categorizes each item by the biological concept and quantitative skill (if any) addressed. Items adapted from the IMCA are indicated in Table 1.

We chose not to include math skill–only items on the outcome assessment, because we anticipated a decrease in student effort had they been asked to complete an overly long assessment for which they were receiving participation credit only.

**We Integrated Quantitative and Biological Concepts in All Aspects of the Course**

We developed course materials over four semesters (Spring 2011 to Fall 2012) in a small experimental section of ~35 students (in Fall 2012, 37 students were enrolled in the course). The section was not advertised as being experimental or as including quantitative content; it was listed alongside other sections with the same course number, and students were allowed to enroll based on their schedule preferences. The number of seats available per class is available to students, so we cannot rule out the possibility that some students chose the course based on its small size (the experimental section had an enrollment cap of 40, as opposed to the typical

---

¹That is, finding the total number of possible sequences of a given length and number of possibilities for each place in the sequence.

²Due to a constraint on the number of items included in the assessment, we did not assess logarithms, exponents, and scale on the outcome assessment. These skills were, however, integrated into course materials.
enrollment cap of 300–350), but according to a postcourse survey, most students chose the experimental section based on schedule constraints over other factors in the Fall 2012 semester.

The guiding principle behind our course design was that the quantitative exercises should support and reinforce the biology curriculum, rather than distract students from it. Applications of quantitative skills and numeracy were treated as part of the material covered: just as the concept “structure affects function” is an idea that students are expected to revisit and apply in various contexts, calculating and making sense of simple probabilities or interpreting graphical representations of data are skills that students practiced many times throughout the course and applied in various situations. We anticipated that applying specific quantitative skills to biology would be foreign to students when first introduced but would be increasingly intuitively applied and understood as the semester progressed, much like more “traditional” biology concepts.

On the basis of initial observations in the pilot sections between Spring 2011 and Spring 2012, we designed new course materials, adapted existing ones, and piloted and refined materials and course design. By Fall 2012, the materials and course had evolved into the form described here. Table 2 indicates the intersection between traditional MCB 181 course topics and quantitative skills in the integrated exercises implemented in the Fall 2012 semester.

Table 1. Pre/post assessment questions by biology concept and quantitative skill

<table>
<thead>
<tr>
<th>Bio concept</th>
<th>Meiosis, genetics, and inheritance</th>
<th>Nucleic acid structure and function</th>
<th>Enzymes, energy, and the reactions of life</th>
<th>Gene regulation and central dogma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math Skill</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>9, 10, 19, 23²</td>
<td>3, 5, 7</td>
<td>1, 21, 2² 22²</td>
<td>13, 17, 24 6</td>
</tr>
<tr>
<td>Algebra and units</td>
<td></td>
<td>4, 8</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Counting and probability</td>
<td>11, 12, 18</td>
<td>2</td>
<td>14, 16</td>
<td>20</td>
</tr>
<tr>
<td>Analyzing graphs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

²Question adapted from the IMCA (Shi et al., 2010).

We Followed a Learner-Centered Approach in Course Design

For each topic covered in the course, we gave students a set of learning outcomes, assignments, and exercises. Before most class meetings, we posted learning outcomes online and assigned students a preclass assignment typically consisting of reading and either online or written questions intended to introduce students to the topic and its vocabulary. In class, lectures were punctuated every 5–20 min with think–pair–share questions (Lyman, 1981, 1987) or in-class written exercises that students completed in groups of three to four. These exercises provided students with guided practice and helped both students and the instructor to assess students’ developing understanding. Students also completed four in-class case studies over the semester. At the end of each week, we assigned an online multiple-choice/multi-select quiz to provide students with additional practice and an opportunity to assess their understanding of the week’s topic. Three quarterly exams (administered every 4 wk) and the cumulative final exam each consisted of multiple-choice and free-response sections and assessed students at the comprehension and application levels.

We integrated quantitative-skill application into course objectives, in-class exercises, quizzes and exams. Typically, we introduced the majority of quantitative-skill application during the in-class practice, although when we suspected many students did not have previous instruction (in particular, in probability and counting), we assigned tutorials accompanied by online instructional videos prior to using the skill in class (online videos are freely available from the Kahn Academy: www.khanacademy.org/video/basic-probability?topic=probability and www.khanacademy.org/video/compound-probability-of-independent-events?topic=probability).

While student groups worked on exercises in class, the instructor and three undergraduate “preceptors” (near-peer instructors selected based on excellent performance in a previous semester’s MCB 181 course and concurrently trained in techniques such as active listening and Socratic questioning) circulated throughout the room facilitating student discussion and providing help where needed. We found that a brief refresher supplied by the instructor, the help of undergraduate preceptors, and/or group work in which students could guide one another in solving problems was sufficient to remind students of the quantitative skills they already possessed and to allow them to complete the problems. This approach was largely successful because students had applied the skills in other contexts previously—they were not grappling with learning entirely new quantitative skills at the same time they were struggling to apply them in novel contexts—and those students who did struggle with the skills themselves had immediate coaching available from other students in their groups, the circulating preceptors, and/or the instructor.

Because each skill was revisited many times throughout the course, students had ample opportunities for building skill proficiency. Also, knowing that they could not simply ride out a few isolated instances of math use during class seemed to encourage students to put more effort into mastering quantitative skills and application than they would have otherwise. It is reasonable to believe that students were less likely to “blow off” quantitative exercises—knowing as they did that the applied skills were going to “come around again” in future exercises and exam questions—than they would have been if quantitative exercises were less frequent, novel to the point of being excessively difficult, independent of explicit course learning outcomes, or not expected to appear on exams.
<table>
<thead>
<tr>
<th>Course Topic</th>
<th>Math Skill</th>
<th>Counting and probability</th>
<th>Reading/Creating graphs and tables</th>
<th>Understanding and manipulating units</th>
<th>Scale, exponents and logarithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analyzing and Engineering Genes</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Photosynthesis</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cellular Respiration and Fermentation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regulating Gene Expression</td>
<td>×</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Translation and the Genetic Code</td>
<td>×</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Translation</td>
<td>×</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Into the Central Dogma</td>
<td>×</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNA Structure, Replication and Repair</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>Meiosis</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mendelian Genetics</td>
<td>×</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell Cycle and Mitosis</td>
<td>×</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell-Cycle Interactions</td>
<td>×</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell Structure and Function</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Membrane Transport</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diffusion and Osmosis</td>
<td>×</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enzyme/Energy of Reactions</td>
<td>×</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Macromolecules</td>
<td>×</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thermodynamics</td>
<td>×</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bonding, Water Chemistry and pH</td>
<td>×</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evolution and Natural Selection</td>
<td>×</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Into the Scientific Method</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Below, we present examples from the course topic “error correction during DNA replication” in order to illustrate our approach. During this unit, quantitative activities emphasized manipulating powers of 10.

**Example of Quantitative Integration throughout Class Materials: Error Correction during DNA Replication**

**Learning Outcomes.** At the beginning of each unit, the instructor posted a list of learning outcomes on the course website. The instructor encouraged the students to print out the learning outcomes, bring them to class, and identify which class activities corresponded with which outcomes. Many students used the lists of learning outcomes as study aids when reviewing for exams. By making desired quantitative learning outcomes explicit and placing them alongside qualitative ones, we emphasized the importance of quantitative skill application in the course and set the expectation that students be able to apply quantitative skills on exams. Such explicit expectations give students incentive to master the quantitative skills in addition to the qualitative material.

**Learning Outcomes—Error Correction during DNA Replication**

- Describe why it is important that cells correct errors that occur during DNA replication.
- Describe DNA polymerase proofreading.
- Explain why DNA is synthesized in the 5' → 3' direction in cells (i.e., what this has to do with proofreading).
- Describe mismatch repair, including how mismatch repair enzymes recognize old and new strands of DNA and why it is important that they do.
- Given error rates in DNA polymerization, DNA polymerase proofreading and mismatch repair, predict the final overall DNA replication error rate (and the number of errors per replication expected in a genome of a given size) a) normally and b) when proofreading or mismatch repair fails.

**Preclass Assignment.** Students were assigned a preclass assignment before the topic was introduced in class. The assignment included reading sections covering DNA replication error correction in a molecular biology text (Biological Science, 4th ed., vol. 1, Scott Freeman) and completing a series of Web activities and tutorials from a commercial online package (Mastering Biology, Pearson Education). The assignment familiarized students with the vocabulary, major enzymatic players, and processes of DNA polymerase proofreading and mismatch repair.

**In-Class Activities.** We opened class with a discussion of mutations—what mutations are and their potential effects on individual cells or organisms and on populations. Students discussed a short list of prompts, first in groups of three to four and then as an entire class. The instructor then delivered a review lecture on DNA polymerase, DNA polymerase proofreading, and mismatch repair. The lecture was punctuated with short discussion questions and the exercises shown in Box 1. Students worked on each exercise in groups of three to four while undergraduate preceptors and the instructor circulated through the classroom monitoring students’ progress and offering help where needed. Giving students time to work through these types of exercises during class (for credit and under the eyes of preceptors and instructors) encouraged them to put forth the necessary effort to grapple with skills they found difficult and provided an opportunity for peer coaching among the students.

**Weekly Quiz Questions.** Online weekly quizzes gave students a chance to apply the knowledge and skills from class on their own. Weekly quizzes served as periodic self-assessment opportunities for the students and provided practice with exam-style questions. They included questions both quantitative and nonquantitative in nature. An example of a quantitative weekly quiz question is shown in Box 2.

---

**Box 1. In-class activity: error correction during DNA replication**

In exercises (a)–(c), students calculated the number of mutations, on average, that a human cell would accumulate every time it divided (a) in the absence of error correction, (b) with DNA polymerase proofreading only, and (c) with both DNA polymerase and mismatch repair. Part (a) is shown as an example.

During the discussion introducing the exercises, the instructor led students in a discussion about when the two error correction mechanisms will come into play—that is, that proofreading will “get a shot at” correcting errors made by DNA polymerase and only the errors that are left over will be corrected by mismatch repair enzymes. They were also told that it is safe to assume for the sake of the exercise that there are not special errors that both correction mechanisms are likely to miss (i.e., it is safe to assume that the error rates for the two are independent).

(a) DNA polymerase III inserts an incorrect base on average in every 100,000 bases added. For a human cell with ~12 billion DNA bases (1 billion = 10^9) in its chromosomes, estimate how many mutations this would result in per cell replication in the absence of error correction. **Please show your work.**

(b) Using the information from (a)–(c), complete the following table:

<table>
<thead>
<tr>
<th>Replication step</th>
<th>Error rate (chance of error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5' → 3' DNA polymerization (DNA polymerase)</td>
<td>1 in 10^5</td>
</tr>
<tr>
<td>3' → 5' DNA polymerase proofreading</td>
<td></td>
</tr>
<tr>
<td>Mismatch repair</td>
<td></td>
</tr>
<tr>
<td><strong>Final overall error rate</strong></td>
<td></td>
</tr>
</tbody>
</table>

(c) If a mutation in the gene coding for DNA polymerase increased its error rate to 1 in 10^5, what would the **final overall** error rate be for DNA replication? **Please show your work.**

(d) *Escherichia coli* cells have an overall mutation rate per cell division that is similar to the one that you found above. Yet populations of *E. coli* evolve much faster than human populations. Why is this?
Box 2. Weekly quiz question: error correction during DNA replication

The three steps that give rise to high-fidelity DNA synthesis

<table>
<thead>
<tr>
<th>Replication step</th>
<th>Error rate (chance of error per nucleotide added)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA polymerization</td>
<td>1 in $10^5$</td>
</tr>
<tr>
<td>DNA polymerase proofreading</td>
<td>1 in $10^2$</td>
</tr>
<tr>
<td>Mismatch repair</td>
<td>1 in $10^2$</td>
</tr>
<tr>
<td><strong>Final overall error rate</strong></td>
<td>1 in $10^9$</td>
</tr>
</tbody>
</table>

If a mutation in the gene coding for DNA polymerase knocked out the DNA polymerase proofreading function but otherwise left the protein functional, the mutant’s **final overall error rate** during DNA synthesis is …

(a) would increase to 1 in $10^{11}$.
(b) would decrease to 1 in $10^{13}$.
(c) would increase to 1 in $10^9$.
(d) would decrease to 1 in $10^9$.
(e) would increase to 1 in $10^7$.
(f) would decrease to 1 in $10^7$.
(g) would increase to 1 in $10^5$.
(h) would decrease to 1 in $10^5$.
(i) would increase to 1 in $10^2$.
(j) would decrease to 1 in $10^2$.

Box 3. Quarterly exam questions: error correction during DNA replication

Questions 5 and 6 refer to the table below.

The three steps that give rise to high-fidelity DNA synthesis

<table>
<thead>
<tr>
<th>Replication step</th>
<th>Error rate (chance of error per nucleotide added)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA polymerization</td>
<td>1 in $10^5$</td>
</tr>
<tr>
<td>DNA polymerase proofreading</td>
<td>1 in $10^2$</td>
</tr>
<tr>
<td>Mismatch repair</td>
<td>1 in $10^2$</td>
</tr>
<tr>
<td><strong>Final overall error rate</strong></td>
<td>1 in $10^9$</td>
</tr>
</tbody>
</table>

5. The dog genome contains $\sim$5 billion ($5 \times 10^9$) nucleotides. If all DNA synthesis error correction enzymes are functioning, roughly ____ mutation(s) will be introduced into the DNA each time a dog cell divides.

(a) less than one
(b) 5
(c) 50
(d) 500
(e) 5000

6. If a mutation increases the error rate of DNA polymerization to 1 in 100, roughly ____ mutations will be introduced into the DNA each time a dog cell divides.

(a) 5
(b) 15
(c) 50
(d) 500
(e) 5000

**Exam Questions.** Similar to the weekly quiz questions, exam questions asked students to apply the (by now) familiar skills practiced in class in similar but (from the perspective of the students) somewhat different contexts. Examples of questions from the quarterly exam, including error correction during DNA replication, are shown in Box 3. Weekly quizzes and exams also included questions asking students to apply skills practiced in one set of contexts to a more novel context—an example of this type of question from the final exam is shown in Box 4. These types of questions, in addition to the reappearance of skills in in-class exercises over many different topics, encouraged generalization and transfer of those skills. For example, the final exam question in Box 4 requires that students apply the skill of interpreting the slope of a line from a scatter plot to a new context—the effect of a breakdown in error correction during DNA replication—and in conjunction with the often-practiced task of using information from a specific table. It is important to note here that, although the example given is of a final exam question, these types of questions were used in in-class activities, weekly quiz questions, quarterly exam questions, and, to a lesser extent, preclass assignments.

**Data Collection and Analysis**

In Fall 2012, we administered the outcome assessment as an online quiz through the Desire2Learn course-management system for each MCB 181R section. We administered the outcome assessment as a precourse test during the first week of classes and as a postcourse test during the last week of classes. Before we asked students to complete the pretest, a study representative visited each classroom and introduced the outcome assessment to students, explaining that the purpose of the study was to measure the students’ growth throughout the semester in order to evaluate the effectiveness of the course and requesting that students answer the questions to the best of their ability without looking up any of the answers or seeking help answering the questions. The representative visited each classroom once again at the end of the semester, reminding students of the purpose of the outcome assessment, asking once again that students try their best without looking up answers or seeking help, and reassuring students that any compensation their instructors had offered for completion (in most cases, a small number of extra credit points) was contingent on full completion of the outcome assessment but not on answering the assessment questions correctly. Students were also reassured that the purpose of the postcourse outcome assessment was to evaluate the effectiveness of the course, not to judge them as students.

Pre- and posttest data were matched for all students who completed both tests and consented to the use of their data.
in the study. Data from students who completed only the pre- or posttest were not included in our analysis. Data from 28 students from the experimental section and 732 students from other sections were included in our analysis. We measured gain as the percentage of possible gain realized: \[ \text{gain} = \frac{\text{post score} - \text{pre score}}{\text{total possible score} - \text{pre score}} \].

In judging significance of multiple-group comparisons, we corrected for the increased possibility of a type I error using the Holm-Bonferroni method (Supplemental Table S1). Table S2 gives Bio, BioMath, and total gains, and SEM for each section.

RESULTS
Incoming Student Performance Supports an Inability to Transfer Skills from Prerequisite Mathematics Courses
Students enrolled in introductory biology at the University of Arizona must demonstrate a certain level of proficiency in pre–college algebra skills through an online math readiness test, Assessment and Learning in Knowledge Spaces (ALEKS; www.aleks.com). Topics assessed by the readiness test include functions, rates of change, linear functions, exponentials, logarithmic functions, and systems of equations. On the basis of our own experience and anecdotal evidence from other introductory biology instructors, however, we questioned whether students were truly proficient at applying the required set of skills. Students’ performance on the precourse outcome assessment supported the belief that students entering MCB 181 were unable to apply quantitative skills from previous courses in the context of biology. Despite having completed mathematics prerequisites that covered the skills tested on the outcome assessment, incoming students scored lower on BioMath items than on Bio items (Table 3; \( p < 0.001 \) for two-tailed \( t \) test comparison between Bio and BioMath precourse scores for all sections combined). This finding is consistent with the literature—it is widely recognized that students have difficulty spontaneously transferring skills to novel contexts (NRC, 2000) and that students in other fields have difficulty transferring even relatively simple mathematical skills to new contexts (Britton, 2002). This result also confirms our earlier observations that students perform significantly lower on tasks requiring application of mathematical skills to biological problems than on context-free use of the same mathematical skills (Supplemental Text ST1 in the Supplemental Materials).

Students in the Experimental Section Had Higher Gains Than Students in Other Sections on BioMath Items
Figure 1 shows the gains on Bio and BioMath items by class section; Figure 2 shows these gains broken down by category. Students in the experimental section had significantly higher BioMath gains than students in each of the comparison sections: 35.97 ± 6.26% for the experimental section versus 18.55 ± 2.94%, 13.89 ± 2.06%, 11.20 ± 2.50%, and 2.36 ± 3.45% for sections 2, 3, 4 and 5, respectively \( (p = 0.020, 0.001, 0.001, \text{and } < 0.001, \text{respectively, when compared} \)
Students in the Experimental Section Did Not Have Lower Gains Than Students in Other Sections on Bio Items

As seen in Figure 1, student Bio gains in the experimental section were not adversely affected by the inclusion of quantitative skills in the course. Bio gains were not significantly different between the experimental section and sections 2, 3, or 4; Bio gains in section 5 were significantly lower than for other sections (see Table S1). This supports the idea that, by integrating quantitative skill application alongside biology concepts, we can increase students’ ability to use mathematics in biological contexts without harming their understanding of the biology concepts.

DISCUSSION

The mathematics skills on which we focused—algebra and manipulating units; scale, exponents, and logarithms; reading and creating graphs and tables; and basic counting and probability—are skills that many (particularly advanced) biology instructors assume their students have and can apply at some level to biological problems. These skills and their application “fly under the radar” when instructors think about what it means to apply mathematics to biological problems: they tend to think instead of more sophisticated skills, such as using differential equations or interpreting probability distributions. Expecting that students can apply these skills (or at least follow their application), instructors present graphical data in lecture to support their course content, ask students to calculate and understand probabilities for basic genetics problems, and otherwise rely on their students’ assumed quantitative skill set. Yet, as seen in our precourse outcome assessment results, students can have a difficult time spontaneously transferring even relatively simple mathematics skills to novel contexts. The resulting mismatch between instructors’ expectations of their students and the abilities that students actually bring into the classroom can be immensely frustrating for both instructors and their students. We have documented here that, with practice and feedback, students can learn to apply quantitative skills to biological contexts. By making quantitative reasoning an explicit objective of our course design, we improved students’ ability to apply their existing mathematical skills to biological contexts.

Integrating Quantitative Thinking Does Not Come at the Cost of Teaching the Biology

One criticism faced by those encouraging math-integration efforts in introductory biology classes is that spending time...
on quantitative skills limits coverage of topical material. This concern has merit—on the face of it, time spent instructing students on reading graphs, algebra skills, and other quantitative concepts is time not spent presenting new biological content. Our results, however, do not support this complaint: students in the experimental section performed no worse than their peers in other sections of MCB 181 on the Bio items on the outcome assessment, while outperforming them on BioMath items. Moreover, our approach minimizes “lost” time by having students practice quantitative skills in the context of the biological material. These results defy the assumption that time spent “doing math” is time spent not doing biology by requiring students to apply quantitative skills and biological concepts in order to solve the problems. This approach is possible, in part, because we emphasize skills that are not entirely new to the students; we are guiding them in how to transfer skills from their required mathematics courses to tackling biological problems. Although problem solving in class does decrease the time available for an instructor to present material, this is time well spent: active learning during class time has been demonstrated to increase students’ class attendance, course performance, and ability to apply concepts; and to decrease the achievement gap among higher- and lower-performing students (Smith et al., 2005; Freeman et al., 2007; Haak et al., 2011). At the same time, many valuable tools—from online companion activities offered alongside textbooks to instructor-recorded lectures—have become available to increase and assess students’ learning before their arrival in the classroom, making prolonged in-class presentation and explanation of the material less necessary.

**Instructional Approach and Course Design Are Crucial When Integrating Mathematics into Biology**

The approach described here is most successfully used with active-learning techniques. Students require a sense of relevance, extensive practice, and prompt feedback in order to develop confidence and proficiency applying quantitative skills in biological contexts. Our approach, however, did not require fundamental changes to the MCB 181 curriculum, and none of the pedagogical strategies we used are revolutionary. We integrated quantitative skill application into the existing curriculum and based our course design on recognized effective teaching practice centered on active learning and providing frequent formative assessment opportunities for the instructor and students. Notably, three of the MCB 181 sections used for comparison (sections 2, 3, and 4) make extensive use of learner-centered techniques: the instructors explicitly define learning outcomes for the students; assign online preclass assignments similar to those assigned in the experimental course; and frequently punctuate their lectures with think-pair-share and/or personal-response system (clicker) questions and written activities mediated by trained undergraduate preceptors. The development of the experimental course was highly informed by these instructors’ methods and experience; one of the three was directly involved in the development of the experimental course and the other two were frequently consulted for advice, material sharing, and feedback on how best to design materials that would be portable into a large lecture-style class in the future. The active-learning environment in these three classes likely explains why Bio gains were not significantly different between any of these sections and the experimental section despite the smaller class size for the experimental course: these three instructors expertly implement techniques that minimize the deleterious effects of large class size on student learning (Smith et al., 2005; Freeman et al., 2007). Bio gains were significantly lower in section 5, a traditionally taught lecture course without an active-learning component, supporting this interpretation of the results.

It is important to note that all of the instructors involved in this study had access to and were involved in giving feedback on the outcome assessment items. Because we shared each section’s pre/postcourse outcome assessment results with the instructors during the instrument’s development, many of the instructors involved in the project gained insights into their students’ strengths, weaknesses, and misconceptions. Instructors reported that they used these results to target specific conceptual areas for course improvement. For instance, instructors reported finding themselves devoting more class time to allowing students to practice interpreting graphs. However, instructors were directed not to use the outcome assessment items in their classes (e.g., as clicker or exam questions).

**Sharing Materials**

One of the greatest challenges of developing the course described was the relative dearth of available classroom materials integrating mathematical tools and molecular and cell biology concepts. We believe that some of the most useful products of this project are shareable class materials that do so. For this reason, we are eager to share any of the materials we have developed with interested instructors. If you are an instructor who would like access to our course materials, please email the corresponding author.

**ACKNOWLEDGMENTS**

We acknowledge the efforts of Athena Ganchorre during the first year developing and teaching the course and developing the first version of the outcome assessment instrument. We thank Susan Jorstad, Bruce Patterson, Ted Weinfurt, Angel Pimentel, and Ramin Yadegari for their feedback on outcome assessment items and their willingness to administer the assessment to their students. Thanks to Carol Bender, Joe Watkins, and the rest of the BioMath community at the University of Arizona for supporting a cross-departmental culture and conversation about integrating mathematics and biology. We also thank Molly Bolger for her input throughout development of the course and outcome assessment, and Katie Southard for her contributions as the teaching assistant and so much more early in course development. Finally, we thank the organizers of the Mountain West Summer Institute at the University of Colorado, Boulder: many of the strategies presented at the institute were instrumental in building our course. This work is supported by a grant to the University of Arizona from the HHMI (52006942).

**REFERENCES**


Integrating Math into an Intro Bio Course


Lyman F (1981). The Responsive Classroom Discussion: The Inclusion of All Students, University of Maryland, College Park: Mainstreaming Digest.


Identification of Threshold Concepts for Biochemistry

Jennifer Loertscher,* David Green,† Jennifer E. Lewis,‡ Sara Lin,* and Vicky Minderhout*

*Department of Chemistry and †Center for Faculty Development, Seattle University, Seattle, WA 98122; ‡Department of Chemistry, University of South Florida, Tampa, FL 33620

Submitted April 9, 2014; Revised May 13, 2014; Accepted May 15, 2014
Monitoring Editor: Mary Lee Ledbetter

Threshold concepts (TCs) are concepts that, when mastered, represent a transformed understanding of a discipline without which the learner cannot progress. We have undertaken a process involving more than 75 faculty members and 50 undergraduate students to identify a working list of TCs for biochemistry. The process of identifying TCs for biochemistry was modeled on extensive work related to TCs across a range of disciplines and included faculty workshops and student interviews. Using an iterative process, we prioritized five concepts on which to focus future development of instructional materials. Broadly defined, the concepts are steady state, biochemical pathway dynamics and regulation, the physical basis of interactions, thermodynamics of macromolecular structure formation, and free energy. The working list presented here is not intended to be exhaustive, but rather is meant to identify a subset of TCs for biochemistry for which instructional and assessment tools for undergraduate biochemistry will be developed.

INTRODUCTION

As a discipline, biochemistry has a set of concepts that most biochemists agree are essential for biochemistry undergraduates to master. For example, the American Society for Biochemistry and Molecular Biology (ASBMB) has recently completed the first phase of an effort to identify foundational concepts and skills for majors in biochemistry and molecular biology (Mattos et al., 2013; Tansey et al., 2013; White et al., 2013; Wright et al., 2013). Another group of biochemists has undertaken an effort to identify “big ideas” in the molecular life sciences as part of the development of the Molecular Life Science Concept Inventory (MLSCI; Howitt et al., 2008; Wright and Hamilton, 2008; see Table 1 for a complete list). Yet biochemistry instructors know from classroom experience that many students fail to master these concepts. A major barrier to achieving mastery is the fact that students bring scientifically inaccurate ideas from their prior experience to biochemistry classrooms, many of which are robust and therefore persistent (Anderson and Schönborn, 2008; Chi, 2008). This problem is further complicated by the fact that these inaccurate ideas may relate to threshold concepts (TCs; Meyer and Land, 2003).

TCs are concepts that, when mastered, represent a transformed and integrative understanding of a discipline without which the learner cannot progress. By focusing on student understanding of TCs in biochemistry, instructors can maximize the impact of classroom instruction toward learning achievement. Meyer and Land (2003), the originators of this educational term, posit that TCs can be identified for any discipline and provide a framework for linking student learning to curricular design. TCs have five defining characteristics:

Transformative: Once a TC is understood, a student’s perception and comprehension of a subject are radically altered. In addition to stimulating cognitive development, learning of TCs can alter a student’s self-perception or sense of identity. For example, students may shift from viewing themselves as students of biochemistry to recognizing that they have begun to think like biochemists.

Irreversible: Once a TC has been deeply understood, students are unlikely to forget it. The concept becomes central to
how students think about everything else in the field. Experts have difficulty remembering how they understood the discipline before deeply understanding TCs.

**Integrative:** TCs bridge concepts within a discipline and among disciplines. Once understood, previously hidden connections within a discipline, and perhaps even across disciplines, are apparent.

**Troublesome:** Most (but not all) TCs are troublesome for students. Concepts can be troublesome for a number of reasons, as described by Perkins (1999). Troublesome knowledge therefore falls into a number of different categories, including ritual knowledge (routine and therefore meaningless), inert knowledge (used only when specifically called on), conceptually difficult knowledge (complex and/or different from personal experience of the world), alien knowledge (conflicts with personal beliefs), tacit knowledge (understanding implicit and therefore often overlooked), and troublesome language (discipline-specific usage of terms; summarized by Meyer and Land, 2006). It is also important to remember, however, that although TCs tend to be troublesome, not all “troublesome knowledge” has a TC at its source.

**Bounded:** Bounded is a fifth category that has been used by educators in some disciplines to identify TCs. This slippery term refers to the fact that TCs sometimes act to define the academic territory of a discipline. The characteristic “bounded” was not used in this study, primarily because biochemistry is by definition interdisciplinary and therefore all concepts in biochemistry lie at the boundaries of academic disciplines (primarily chemistry and biology).

According to Ross and colleagues, who have studied TCs in biology, “we need to emphasize that while academics and teachers identify the content knowledge as troublesome or problematic, the threshold concepts which underlie the problematic difficult content knowledge receive the least attention in teaching” (Ross et al., 2010, p. 170). Therefore, the primary purpose in identifying TCs for a discipline is to provide a starting point for focused curricular redesign, because an intentional approach to teaching TCs is likely to result in the greatest improvement in student learning (Entwistle, 2008; Perkins, 2008). Owing to the importance of TCs in mastery of a discipline, it is reasonable to expect that, if more time were spent developing student understanding of them, the learning of additional core concepts would happen more readily, deepening student understanding of the discipline. Land and colleagues suggest that TCs “be viewed as the ‘jewels in the curriculum’ insomuch as they can serve to identify crucial points in the curriculum that provide opportunities for students to gain important conceptual understandings” (Land et al., 2006, p. 198).

TCs research is an iterative process involving disciplinary experts, students, and education experts. According to Cousin, research into TCs “offers an analytical framework for bringing into view conceptual and/or affective difficulties in the disciplines” (Cousin, 2009, p. 201). TCs research differs from other modes of qualitative research in that it places disciplinary experts at the center of the research endeavor. Furthermore, this approach places teachers and students in dialogue with the ultimate goal of improving teaching and learning. Therefore, according to Cousin, “research and pedagogy overlap” in this process of transactional curriculum research (Cousin, 2008, p. 269).

Since the term was coined in 2002, TCs have been the subject of study across all major branches of higher education, with specific studies, for example, in biology (Taylor, 2008; Kinchin, 2010; Ross et al., 2010) and the health sciences (Clouder, 2005), including dentistry (Kinchin et al., 2011), occupational therapy (Rodger and Turpin, 2011), and nursing (Stacey and Stickley, 2012). Work has begun on identifying TCs in chemistry (Park and Light, 2009). As yet, however, TCs essential to biochemistry have not been characterized.

Work on TCs in biology provides a starting point for thinking about possible TCs in biochemistry. For example, submicroscopic events, energy transformations, and the ability to think correctly about scale are biology TCs that may also be relevant for biochemistry. Additionally, as part of their work to create the Biology Concept Inventory, Garvin-Doxas and Klymkowsky identified the role of randomness in processes such as diffusion and evolution as particularly troublesome for students (Garvin-Doxas and Klymkowsky, 2008). They suggest that a deep understanding of the role of emergent processes in biology and chemistry is fundamental and that, once students understand these processes, their view of the discipline changes dramatically. Investigation of these concepts in a biochemistry context may indicate that proper understanding of them is also transformational for students learning biochemistry.

In this study, we investigated two major research questions:

1. What potential TCs for biochemistry emerge from discussions with groups of faculty and students?
2. What knowledge statements related to each TC can be generated using an iterative process that considers both expert and student perspectives? Using TCs research in other disciplines as a starting point, we undertook a five-phase, national effort to engage faculty members and undergraduates in a process to identify and refine TCs for biochemistry. This study is part of a larger, systematic approach to improve student learning in undergraduate biochemistry courses nationwide through development of instructional and assessment materials focused on TCs. Therefore, the goal was not to produce an exhaustive list of TCs for biochemistry, but rather to identify a manageable number to move forward into the curriculum-development phase of the project. Finally, this study can act as a model for others wishing to investigate TCs in their area of interest.

### Table 1. Previously published concepts for biochemistry

<table>
<thead>
<tr>
<th>Concept Area</th>
<th>Relevant Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASBMB foundational concepts</td>
<td>Evolution, matter and energy transformation, homeostasis, biological information, macromolecular structure and function</td>
</tr>
<tr>
<td>Molecular life sciences</td>
<td>Molecular evolution, self-assembly, compartmentalization, information, and communication, regulation, catalysis, energy and organization, complexity of molecular structures, and the aqueous environment</td>
</tr>
</tbody>
</table>

*a* Tansey et al., 2013.  
*b* Wright and Hamilton, 2008.
METHODS

Project Context
This work is part of a National Science Foundation (NSF)-funded project to improve student understanding of TCs in biochemistry so instructors can maximize the impact of classroom instruction on student success. Because biochemistry is an interdisciplinary field taught in biology, chemistry, and medical departments, the project involves a diverse national community of experts to 1) identify and confirm TCs that are critical for learning in biochemistry, 2) develop assessment instruments to measure student achievement related to identified TCs, 3) design classroom activities to target TCs, 4) support a community of innovative biochemistry educators, and 5) disseminate classroom activities and assessment tools. The process for identifying TCs, Chapter 12: Transactional Curriculum Inquiry: Researching Threshold Concepts described below, is based on work by Cousin and others and involves input from both faculty experts and students (Cousin, 2009, p. 201–212).

Faculty Participants
More than 70 faculty members were involved in this process, primarily as participants in one or two of three workshops held in the Summer of 2013. The first was a 2-d interdisciplinary life sciences workshop held at the University of Minnesota in June 2013. Twenty faculty members with significant teaching experience and diverse expertise in the molecular life sciences attended. Some attendees were textbook authors and/or education researchers, and several had been involved in the ASBMB and MLSCI projects described above. This workshop was cofacilitated by four of the authors. The second was a 3.5-d biochemistry core collaborators workshop held at Seattle University in August 2013. Nineteen faculty members with biochemistry teaching experience and diverse expertise in the molecular life sciences attended. Some attendees were textbook authors and/or education researchers, and several had been involved in the ASBMB and MLSCI projects described above. This workshop was cofacilitated by the two authors who are biochemists (J.L. and V.M.). The third was a 2.5-h dissemination workshop to introduce faculty to the idea of TCs and to solicit further feedback on potential biochemistry TCs, held as part of the ASBMB Symposium on Student-Centered Education in the Molecular Life Sciences at Seattle University in August 2013. The 38 attendees included high school teachers and faculty from a variety of colleges and universities. This workshop was cofacilitated by the two authors who are biochemists (J.L. and V.M.).

Ensuring diverse participation was a priority, especially for the interdisciplinary life sciences and biochemistry core collaborators workshops, where substantial work of the project was accomplished. Participation at these two workshops was by invitation, and attendees were recruited from a variety of institution types and departments. Faculty members from small colleges comprised the largest percentage of participants (44%), but faculty members from large research universities were also well represented (33%). Faculty members from master’s-level universities (15%) and 2-yr colleges (8%) comprised the remainder of the participants. Gender distribution was reasonable, with 36% of participants being men and 64% being women. An effort was made to promote ethnic diversity among workshop participants and to include members from minority-serving institutions. Of the 13 faculty members representing these diversity characteristics who were invited, however, only four were able to attend either of the workshops. Diversity statistics were not compiled for the dissemination workshop, because it was organized in conjunction with an ASBMB event.

Student Interviews
Focus group interviews probing student understanding of potential TCs were conducted at five geographically dispersed institutions in Spring semester and Fall semester 2013. Two institutions were private, and three were public. One institution was a doctoral/research university, three were master’s-granting universities, and one was a 2-yr college. Of the five participating institutions, two were minority-serving (one Hispanic-serving university and one historically black university). Students at four of the five interview sites were enrolled in biochemistry at the time of the interview or had recently completed a junior/senior undergraduate biochemistry course. Students at the remaining institution were enrolled in organic chemistry. We opted to include one group of students who had not yet enrolled in biochemistry based on a suggestion by two community college instructors who attended the interdisciplinary life sciences workshop. They asserted, and we agreed, that it is important to investigate student understanding of these concepts in prerequisite courses. This particular group of organic chemistry students was chosen because their instructor teaches a number of organic chemistry concepts using a biochemistry context. In compliance with human subjects protocols approved by all participating institutions, no further demographic data were collected about students.

A total of nine 1-h and three half-hour focus groups were held. Three to eight students participated in each interview, for a total of 56 students (10 in phase 1 and 46 in phase 2). Interviews at four of the five institutions were conducted in pairs by four researchers (J.L., D.G., S.L., and V.M.). Interviews at one institution were conducted by a single researcher (J.L.). The interviews followed a semistructured format (see the Supplemental Material) and were recorded and transcribed. Thematic analysis was performed on interview recordings using the following process. Immediately following interviews at each institution, the pairs of researchers who conducted the interviews debriefed on the experience and discussed themes that arose. Throughout the semester that interviews were being conducted, all four researchers (two biochemistry educators, a faculty developer with significant experience analyzing interview data, and a biochemistry student) met on an ongoing basis and engaged in iterative cycles of discussion related to major and minor themes. One researcher acted as the lead coder, listening to and taking notes on all interviews. As the semester continued, researchers also read the notes generated by the lead coder. The faculty developer and student played important roles in asking clarifying questions, especially those related to participant affect and ways in which students may have interpreted questions differently than expected, as a consensus emerged through discussion. No major disagreements arose.

A summary of findings with illustrative quotes from the interviews is presented in the results section of this document. All data presented are labeled with both a focus group number and a letter designating different individuals within a focus group. For example, Student G15-A is one student (Student A) from focus group 15 (G15).
Biochemistry Threshold Concepts

Table 2. Summary of project phases

<table>
<thead>
<tr>
<th>Phase</th>
<th>Participants</th>
<th>Goal(s)</th>
<th>Outcome(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Pilot student focus group interviews</td>
<td>10 students from one institution</td>
<td>Test focus group protocol on a small scale; Investigate student understanding of a potential threshold concept (equilibrium)</td>
<td>Refined approach to writing semistructured focus group questions; Insights about how visual representations can present barriers to student comprehension</td>
</tr>
<tr>
<td>2: Interdisciplinary life sciences workshop</td>
<td>20 experts in biology, biochemistry, and chemistry education</td>
<td>Generate a draft list of threshold concepts for biochemistry</td>
<td>List of provisional threshold concepts</td>
</tr>
<tr>
<td>3: Biochemistry core collaborators workshop and dissemination workshop</td>
<td>19 biochemistry teaching experts; 38 high school, college, and university educators in the molecular life sciences</td>
<td>Revise draft list of threshold concepts for biochemistry; Begin planning for student interviews</td>
<td>Refined list of provisional threshold concepts; Draft protocols for student interviews</td>
</tr>
<tr>
<td>4: Student focus group interviews</td>
<td>46 students from five different institutions</td>
<td>Explore student understanding of three provisional threshold concepts (equilibrium; interactions; individual versus populations of molecules)</td>
<td>Insights about students’ incorrect ideas related to potential threshold concepts; Deeper understanding of specific concepts that pose barriers to student learning</td>
</tr>
<tr>
<td>5: Data analysis and determination of a working list of threshold concepts</td>
<td>Authors and biochemistry teaching experts</td>
<td>Produce a working list of threshold concepts using an iterative process of data analysis and feedback from experts</td>
<td>Working list of five threshold concepts to form the foundation for development of instructional and assessment materials</td>
</tr>
</tbody>
</table>

The a and b in columns 1 and 2 link workshop in column 1 to participants in column 2.

RESULTS

Our process for identifying TCs was composed of five phases (Table 2). The process and outcomes for each phase are described in considerable detail in the following sections. Our process warrants a rich description for two reasons. First, TCs research relies on dialogue among participating experts and students (Cousin, 2008), and the research process is therefore inseparable from the results. Second, a more detailed explanation of this complex process enables others to replicate it, so as to identify TCs in their own disciplines or to further explore additional TCs in biochemistry.

Phase 1: Pilot Student Focus Group Interviews

Before faculty workshops in Summer 2013, two focus group interviews were conducted with 10 students from one institution. All students were enrolled in biochemistry or had recently completed a biochemistry course at the time of the interviews. The purpose of these interviews was twofold. First, we wanted to pilot focus group interview protocols on a small scale before undertaking large-scale interviews. More importantly, we aimed to explore student understanding of a potential TC before conducting faculty workshops. Starting the TCs research process by talking with students instead of faculty is not the approach that has been typically used in other disciplines (G. Cousin, personal communication). However, because our perspective is informed by our student-centered teaching practices, using student responses as formative feedback to guide our curriculum design and research process is natural to us. Reflecting on our process, we realize that starting the dialogue with students allowed us to keep conversations with faculty infused with a student perspective. This is important, because the irreversible nature of threshold concepts means that experts often forget how the discipline appeared to them when they were students (Meyer and Land, 2006).

The concept investigated in pilot focus groups was equilibrium, because it had already been identified as a TC in biology (Ross et al., 2010) and was therefore also a strong candidate TC for biochemistry. Questions centered around examination of two figures: 1) a schematic showing steroid binding to a receptor and subsequent activation of transcription and 2) chemical reactions showing multiple linked equilibria (see the Supplemental Material). Analysis of student responses revealed two useful generalizations. First, students were confronted with the fact that biochemical schematics often obscure underlying chemical principles. Specifically, the steroid image seemed to communicate that steroid binding and transcriptional activation are unidirectional events that occur in simple sequential manner. Some students recognized that many of the processes depicted could be reversible, but others struggled to recognize that relative concentrations of molecules determine the direction of pathway flux at any given moment. Consultation with expert biochemists revealed that although these details are not explicitly depicted in schematic images, they are “seen” by experts who rely on tacit knowledge of cellular conditions and chemical principles. Second, we observed that all students struggled with the second scenario, in which multiple connected reactions were presented.
Reflection on our part led to the conclusion that we were not able to effectively construct a meaningful dialogue with students regarding these questions. Consequently, we recognized the importance of trying to anticipate what might sidetrack or confuse students when writing semistructured interview protocols for focus groups. Finally, data collected from these interviews were used to inform the faculty workshops described in the sections on phases 2 and 3. Specifically, we recognized the importance of using exploration of images as a means to investigate student and expert understanding of a concept.

**Phase 2: Interdisciplinary Life Sciences Workshop**

The goal of the interdisciplinary life sciences workshop was to draw on the expertise of a diverse group of educators to develop a draft list of TCs for biochemistry. Workshop sessions were designed to provide enough structure to ensure productivity but were left sufficiently open to allow for creative contributions. Facilitators were careful not to influence the conversation with preconceived ideas, which position was supported by the presence of a non-scientist on the facilitation team, and made an earnest effort to allow every voice to be heard. Over the 2 d of the workshop, a draft list of 15 TCs (Table 3, first column) was generated through the four major activities described in the following sections.

**Introduction to TCs.** Participants were introduced to the idea of TCs by one of the authors (not a biochemist, but a faculty developer, D.G.) through wrestling with “signification,” a known TC from literary studies (Meyer and Land, 2003). Through this activity, faculty members experienced the idea of TCs from a student perspective, because none of them had expert understanding of the term “signification.” This term subsequently became a useful reference point for clarifying participants’ definitions of terms and phrases.

**Identification of TCs Using the Defined Characteristics.** Participants brainstormed “troublesome” concepts in biochemistry individually and then discussed and refined their ideas in small groups. After reporting out small-group ideas and engaging in a whole-group discussion, individuals chose their top 10. The workshop facilitators compiled the individual top 10 lists over lunch, producing a list of 32 troublesome concepts. Analysis of the list of 32 troublesome concepts revealed a subset of 10 concepts that received a large number of votes. Next, small groups focused on one further characteristic of TCs (either irreversible, integrative, or transformative) and determined whether the 10 troublesome concepts also met their new criterion. Group members then rotated to other groups using a jigsaw approach in order to share ideas on the three additional TC characteristics and receive feedback. At the end of this activity, all 10 concepts had been determined to meet all four characteristics. The decision to start with “troublesome” before refining the list was informed, as mentioned previously, by wanting to activate participants’ knowledge of the student perspective (as informed by teaching experiences) at the beginning of the process.

**Image Analysis.** A separate activity was introduced as a way of triangulating the previous findings by approaching biochemistry from a different perspective: that of visual representation. This decision arose out of our pilot focus groups, where we discovered images presented their own barriers to student learning and comprehension. Prototypical biochemical images were identified by the authors in collaboration with expert biochemistry educators before the workshop. The identified images related to a steroid hormone signaling pathway, the citric acid cycle, a molecular view of DNA polymerization, free-energy changes associated with glycolysis, a molecular view of hemoglobin binding 2,3-bisphosphoglycerate, and a molecular view of an enzyme–substrate interaction. Participants analyzed these images in

---

**Table 3. Threshold concepts resulting from project phases**

<table>
<thead>
<tr>
<th>Phase 2</th>
<th>Phase 3</th>
<th>Phases 4 and 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual versus populations of molecules</td>
<td>Individual versus populations of molecules</td>
<td>Steady state</td>
</tr>
<tr>
<td>Energy transfer and transformation</td>
<td>Energy transfer and transformation</td>
<td>Biochemical pathway dynamics and regulation</td>
</tr>
<tr>
<td>Reversibility of processes/reactions (equilibrium)</td>
<td>Equilibrium, including ideas related to steady state, open systems, reversibility of processes, and dynamic processes</td>
<td>Physical basis of interactions</td>
</tr>
<tr>
<td>Steady-state and open systems</td>
<td>Intermolecular and intramolecular interactions (how molecules interact)</td>
<td>Thermodynamics of macromolecular structure formation</td>
</tr>
<tr>
<td>Intermolecular and intramolecular interactions</td>
<td>Spatial/scale relationships</td>
<td>Free energy</td>
</tr>
<tr>
<td>Spatial/scale relationships</td>
<td>Spatial/scale relationships</td>
<td></td>
</tr>
<tr>
<td>Entropy/enthalpy/free energy</td>
<td>Entropy/enthalpy/free energy</td>
<td></td>
</tr>
<tr>
<td>pH, pKₐ, and charge</td>
<td>pH, pKₐ, and charge</td>
<td></td>
</tr>
<tr>
<td>Moving between different structural</td>
<td>Biochemical visualization</td>
<td></td>
</tr>
<tr>
<td>representations of molecules</td>
<td>New concept added in phase 3: regulation</td>
<td></td>
</tr>
<tr>
<td>Reading and understanding chemical formulas and structures</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interpreting graphs</td>
<td>Biochemical visualization</td>
<td></td>
</tr>
<tr>
<td>Mental models and text representations</td>
<td>Probabilistic thinking</td>
<td></td>
</tr>
<tr>
<td>Biochemical visualization</td>
<td>Dynamic/fluctuating processes</td>
<td></td>
</tr>
<tr>
<td>Randomness versus directedness</td>
<td>Randomness versus directedness</td>
<td></td>
</tr>
</tbody>
</table>
small groups to uncover the concepts embedded within each image that are obvious to experts but may be hidden to students. Each group was also asked to identify three to seven concepts or skills that an individual would need to have mastered in order to deeply understand the image. The concepts and skills identified were compiled and analyzed to determine whether they met all definitions for TCs. At the end of this process, 12 concepts had been determined to be potential TCs. Some of these overlapped with concepts identified in the above described process, but others were new.

"Grain-Size" Activity. A combined list of concepts from the two above-mentioned activities was evaluated for “grain size.” Participants were asked to determine whether the concepts were too broad, too narrow, or an appropriate size to be the subject of new instructional and assessment tools. Fifteen concepts were determined to be “just right.” These are given in Table 3, column 1.

Phase 3: Biochemistry Core Collaborators Workshop and Dissemination Workshop
The major goal of both the biochemistry core collaborators workshop and the dissemination workshop was to investigate and refine the TCs list generated at the first workshop using the lens of biochemistry classroom teaching experience. Unlike attendees at the interdisciplinary life sciences workshop in phase 2, the primary teaching responsibility of the majority of participants at these workshops was biochemistry.

After one and a half days devoted to the following activities at the biochemistry core collaborators workshop, a refined list of nine potential TCs (Table 3, second column) was produced.

Introduction to TCs. Participants were introduced to the idea of TCs through wrestling with the concept of “signification” (as explained earlier).

Refinement of TCs Using the Defined Characteristics. Participants performed deep analysis of the draft TCs list to determine whether the items met all of the criteria of TCs (troublesome, transformative, integrative, irreversible).

Analysis of Additional TCs. Participants were given the opportunity to identify additional TC candidates. Small groups worked to determine whether these newly proposed TCs met all four criteria. At this point, regulation was added as a TC.

Ranking of Potential TCs. The process for choosing the final set of TCs focused on meeting three conditions: 1) Concepts are TCs for biochemistry. 2) Concepts relate to several foundational concepts in biochemistry. 3) Authors and faculty collaborators could envision developing effective assessment and classroom materials. Furthermore, because our ultimate goal in identifying TCs is to develop instructional and assessment tools, we needed a mechanism to prioritize the draft list of concepts in order to identify the three to five concepts to move forward into curriculum development. Participants therefore voted for their top five concepts in terms of perception of utility within their own teaching. Five concepts clearly rose to the top (given here in rank order): equilibrium, inter- and intramolecular interactions, pH and pKₐ, regulation, and visualization.

To gain a broader perspective, participants at the much shorter dissemination workshop were also asked to perform the same ranking. After an orientation to the characteristics of TCs and an opportunity to comment on the existing list, the top five TCs from this workshop were (given in rank order) equilibrium, entropy/enthalpy/free energy, energy transfer, individual versus populations of molecules, and inter- and intramolecular interactions.

Phase 4: Student Focus Group Interviews
TC research relies on input from experts and students. Therefore, student focus group interviews were conducted in order to determine 1) whether concepts identified by faculty are indeed troublesome for students and 2) what, specifically, students think about each of the concepts. Findings related to the second point were used to refine the TCs and to define knowledge statements associated with each concept.

The first step in preparing for the focus groups was to determine which concepts to further explore with students, since nine was too many given the scope and goals of our project. Student understanding of candidate concepts was garnered from a variety of sources including 1) preliminary interviews; 2) the literature on learning in chemistry, biology, and biochemistry; 3) analysis of MLS CI data (T. Wright, personal communication); and 4) personal teaching experience. All of these factors were considered when deciding which concepts to explore using focus groups. The two authors who are biochemists (J.L. and V.M.) decided that equilibrium, intra- and intermolecular interactions, and individual versus populations of molecules would be the topics discussed in student focus groups. Interview protocols were developed based on known student alternate conceptions related to these concepts and on expert understanding of these concepts received from six participants in the interdisciplinary life sciences workshop (see the Supplemental Material).

As described in Methods, focus group interviews were conducted at five diverse institutions with a total of 46 students. Students from four institutions were enrolled in biochemistry at the time of the interviews, and students from one institution were enrolled in organic chemistry. As expected, wide variability in student understanding was observed, but a number of commonalities were also noted. Quotations are given to support key ideas summarized below when possible. Owing to the fact that focus group interviews led to dialogue among participants, it was sometimes challenging to capture brief snapshots of the conversations to use as examples.

In general, students tended to rely on contexts and concepts that were the focus of course work at the time of the interview—perhaps a matter of overzealous transfer (Schwartz et al., 2012). For example, one class was just starting discussion of enzyme kinetics and therefore used enzyme kinetics as a context for many of their responses. Another group was in the midst of studying glycolysis and therefore discussed pathway flux and the role of rate-limiting steps in metabolism, concepts that students at other institutions did not mention. Yet despite the variety of perspectives, a number of common trends emerged. All groups discussed the complexity of biochemistry compared with systems studied in general and organic chemistry. Related to this, some students implied that the rules of chemistry used in previous course work are different or do not apply to chemistry in biological systems. Students also commented on the depth of understanding they needed to succeed in biochemistry; they acknowledged that memorization does not suffice in studying.

Vol. 13, Highlights of 2014
biochemistry. Finally, students described the difficulties they encountered when moving from the molecular visualizations used in general and organic chemistry to new, often much more abstract, representations of biological macromolecules. Common incorrect or incomplete ideas related to each of the three targeted concepts are described in greater detail below.

**Equilibrium.** A major incorrect idea related to equilibrium was that biological systems are at equilibrium. This idea seemed to stem largely from an everyday use of the term *equilibrium* to mean “balanced” or “just right.” When asked what comes to mind when they hear the term equilibrium, one student (Student G28-A) said the following: “Everything is happy, yeah that’s the easiest way to say it.” When asked whether reactions in the body are at equilibrium, a student in another focus group (Student G30-A) described equilibrium as meaning “normal”: “I think it depends, you know, if you are sick . . . or if . . . I mean, like your body should be normal at equilibrium.” Students also described an intuitive sense that concentrations of molecules are properly maintained within the body, and, therefore, they concluded that biological systems must be in equilibrium. An example is given here:

Interviewer: Are reactions in the body at equilibrium?

Student G13-A: I think it has to be at equilibrium. Like she said, if there’s too much of something, it’s gonna throw, you know, one of the systems off.

Another issue related to complexity and the ways in which biochemical systems appear to differ from systems studied in general chemistry. Students could describe the fact that multiple reactions in a pathway are required for net production of a given metabolite, but they could not reconcile net production with the idea of reactions going back and forth as they had learned when studying equilibrium in previous courses. This problem was complicated by the fact that students believed that biological reactions are in equilibrium. For example, one student (Student G21-A) discussed dissonance between prior ideas about equilibrium and net production of a substance in the body: “I think it’s hard to grasp the concept of something that can be in equilibrium going back and forth, but also having to favor one side. Like, equilibrium, it has the word equal in it, so in my mind, they should be the same. So it’s hard to have dynamic flow of something to favor something when they’re in equilibrium. So it’s just a difficult concept to grasp.”

We also observed instances in which students concluded that the chemical rules related to equilibrium must apply differently, or even not apply at all to biological systems: Student G31-A said, “I think like it’s not hard to understand equilibrium, but there’s a key difference when you’re looking at, like when you’re studying equilibrium in gen chem or analytical chem versus in biochem.” Finally, in cases in which students lacked a foundational understanding of chemical equilibrium, equilibrium constants, and Le Chatelier’s principle, they were unable to apply these ideas to biochemistry.

**Intra- and Intermolecular Interactions.** Focus group interviews revealed that students had superficial, memorized, or incorrect understanding of the physical basis of noncovalent interactions such as hydrogen bonds, dipole–dipole interactions, and van der Waals interactions. Students could name the interactions, and some could discuss the role of polarizable electron clouds in these interactions, but they struggled to make generalizations about the electrostatic basis of the interactions. For example, when asked to list intermolecular interactions, several students in one focus group named van der Waals interactions, but when asked to explain the basis for van der Waals interactions, students struggled, as illustrated here:

Student G14-A: Those are the ones that come in close contact with each other. They don’t necessarily bond in a sense. They interact closely. I can show you a picture of that too.

Interviewer: What is the basis of their interaction, why are they interacting?

Student G14-A: Close proximity.

Interviewer: Close proximity?

Student G14-B: But what attracts them?

Student G14-C: I think it’s electronegativity. I don’t want to . . . I think I might be wrong.

In another example, students again named van der Waals interactions as a noncovalent interaction, but then were unable to provide any detail about the physical basis of these interactions or how they would appear in biochemistry:

Student G16-A: I’d just say like in my classes—I don’t know about what you all had—but we never really talked that much about van der Waals interactions. Other than the fact that they said, you know, everybody has them and they keep it from floating away.

Student G16-B: Yeah, I think I heard at one point in time that it just adds to stabilization, to, well, especially to protein structure. Yeah, it’s not really touched on.

Students also did not feel confident in making predictions about interactions in complex macromolecules. For example, when shown an image of a protein, one student named two different types of interactions, but when asked for specifics, he was uncertain:

Student G14-D: I said there was hydrogen, well I guess not, never mind. I said ionic interactions on the protein surface.

Interviewer: Which part of the protein surface?

Student G14-D: I’m not sure. I guess anywhere on the surface.

In a different focus group, one student (Student G20-A) noted that molecules are much more complex in biochemistry as compared with those studied in prerequisite courses: “So in gen chem it is just thinking about molecules separately, not the overall picture of, like, if we kept folding this in on itself again and again.” Furthermore, when we did hear students making the statement that hydrophobic amino acid residues move to the inside of folded proteins and hydrophilic residues remain on the outside, no further elaboration for this phenomenon could be produced. When asked to explain, for example, students claimed that hydrophobic groups want to be on the inside of the protein, erroneously attributing intentionality to proteins: Student G20-B said, “So polar things want to interact with polar things and the non-polar things don’t want to interact with the polar so they are going to hide somewhere else.” The fact that student understanding of the enthalpic and entropic factors affecting an
event like protein folding was limited to what appeared to be memorized statements strongly suggests that more needs to be done to develop deep conceptual understanding of the energetics that drive interactions and folding of macromolecules in an aqueous environment.

**Individual versus Populations of Molecules.** Examination of student responses related to the concept of individual versus populations of molecules revealed no uniform TC. Students acknowledged that they had not previously considered the term “population” in a biochemical context. For example, during one focus group, students discussed the fact that they had never previously considered that there is more than one molecule of hemoglobin in each red blood cell:

- Student G29-A: I was unaware that there were so many hemoglobin molecules in a red blood cell. I wasn’t aware of that for the longest time until I got to this class.

- Student G29-B: Did you think there is just one?

- Student G29-A: I thought there was only a certain number of hemoglobin molecules per red blood cell. [Whispers] But there’s millions.

- Student G29-B: I’m learning this right now [as we speak].

- Student G29-C: Honestly, when I thought of a red blood cell, I saw pictures of how big they were—but I honestly, I thought, mechanism-wise, I thought it was one hemoglobin [until this conversation].

- Student G29-A: Yeah, ‘cause I was reading about sickle cell anemia and how they can form these polymerizing rods of hemoglobin and so I thought, there must be lots. So I looked it up on Google. But no one ever specifically said, like, there’s a lot of hemoglobin, this is just one small chemical.

- Student G29-B: I can see—if you think about—hemoglobin is this, like, protein—a proteinous thing. And proteins are small compared to a cell. So it makes sense, but I just never made that association before.

Furthermore, students sometimes struggled to move back and forth between representations of individual molecules and measurable properties arising from average characteristics of a population of molecules. However, while we recognized that these ideas are transformative for students, we found that understanding these concepts, once presented, did not seem particularly troublesome or integrative for students. Instead of retaining this concept as a TC, we decided to add explicit treatment of individual versus populations of molecules as a guiding principle in our future instructional materials design process.

**Phase 5: Data Analysis and Determination of a Working List of Threshold Concepts**

Ideas collected from disciplinary experts and students were compiled and analyzed in order to identify the specific concepts that are troublesome for students and to discern how deep understanding of these concepts leads to a transformed and integrated understanding of biochemistry. Using an iterative analytic process, described in greater detail below, the authors worked in collaboration with project advisory board members and participants from the interdisciplinary life sciences workshop to produce a detailed description of five TCs that will be used as foundation for future development of instructional and assessment materials.

Careful analysis of all student focus group data was essential in developing the finalized list of TCs. Interestingly, the three concepts used for the interviews (equilibrium, intra- and intermolecular interactions, and individual versus populations of molecules) appear to be significantly different from the finalized list given in Table 4. Yet all of the ideas described in Table 4 originated from analysis of student conceptions and comparison of student perspectives with those of experts. Equilibrium serves as a good illustration. Throughout the faculty workshops, equilibrium appeared as a relevant concept, but the term elicited different meanings for different people. Some thought of equilibrium constants and Le Chatelier’s principle, others thought of dynamic processes in general, still others focused on the fact that biological systems are not in a state of chemical equilibrium. Going into the interviews, we were interested in discovering from students what exactly about equilibrium was problematic or important. As mentioned previously, talking with students revealed several common issues. First, students recognize that biological systems are much more complex than systems they studied in general chemistry, but instead of considering how to apply equilibrium concepts to the more complex system, many concluded that the chemical “rules” must be different for biochemistry. Second, when asked whether biological systems are at equilibrium, most students relied on everyday usages of the term to mean “just right,” “stable,” or “balanced.” As a result, they concluded that biological systems must be in equilibrium. This colloquial use of the term *equilibrium* had not been apparent to us previously, but after hearing it from students, we were able to see how the misuse of language could be the root of alternate conceptions that we had previously recognized. Finally, related to the issue of terminology, students were upfront about the fact that they did not know what “steady state” means in a biological context. When we compared these student conceptions with the expert understanding we had collected before interviews, a new vision of the TCs related to equilibrium began to emerge. Similar to students, experts indicated that the concept we had called “equilibrium” actually contained a number of different concepts, including biological steady state, reversibility of chemical and physical phenomena, and the complexity of multiple, interconnected equilibria commonly observed in biological systems. As a result, two concepts—steady state and biochemical pathway dynamics and regulation—were ultimately determined to be TCs (Table 4). A process similar to that described for equilibrium was used with other initial concepts to arrive at all of the concepts shown in Table 4.

After deciding on the five major concepts, we next sought to define the specific characteristics that make each of the five identified concepts TCs. To accomplish this task, we sought to clarify how or why understanding a given concept is troublesome, transformative, integrative, and irreversible. The first step in this process was to define knowledge statements that present a more detailed picture of each TC (Table 4). We started the process of producing knowledge statements by systematically analyzing the interview data for incorrect or incomplete ideas held by students. Such ideas were considered to be troublesome. We then generated the knowledge
Table 4. Refined threshold concepts and knowledge statements

<table>
<thead>
<tr>
<th>Name</th>
<th>Knowledge statement(s)</th>
<th>Biochemical ideas that are unlocked once this concept is understood</th>
<th>Connections that were invisible before deep understanding of the concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steady state</td>
<td>Living organisms constitute open systems, which constantly exchange matter and energy with their surroundings, yet net concentrations remain relatively constant over time. This dynamic, yet outwardly stable condition is referred to as a steady state. “Steady” is not synonymous with chemically “stable.” Concentrations are determined by kinetic, rather than thermodynamic, factors. Hence, biological systems do not exist in a state of chemical equilibrium. If an organism reaches chemical equilibrium, its life ceases. Consequently, organisms have evolved extensive regulatory systems for maintaining steady-state conditions.</td>
<td>Steady state is an emergent process that results from regulation of numerous biological reactions. Steady state is a metastable condition that can be maintained only because of constant input of energy from the environment. Steady state defines the conditions of life under which chemical reactions take place in cells and organisms. Therefore an understanding of steady state is necessary in order to correctly contextualize all of biochemistry.</td>
<td>Once the condition of steady state is recognized, the purpose of complex regulatory systems in maintaining steady state and their connections to each other become apparent. Once the metastable nature of steady state is recognized, the importance of multi-tiered energy storage systems (starch, glycogen, triglycerides, etc.) becomes apparent.</td>
</tr>
<tr>
<td>Biochemical pathway dynamics and regulation</td>
<td>Reactions and interactions in biological systems are dynamic and reversible. Directionality of processes depends on the free energy and relative concentrations of reactants and products available. Observable flux is the net result of forward and reverse processes. Enzymes control rates of forward and reverse reactions. Enzyme activity is highly regulated.</td>
<td>Chemical drivers result in bulk (emergent) properties observed in biological systems. Enzyme-mediated regulatory mechanisms allow pathways to be sensitive and responsive to the needs of the organism. Enzymes act as gatekeepers rather than drivers of chemical change.</td>
<td>Once these concepts are understood, predictions can be made about 1) how biochemical pathways are likely to respond to changes in environmental conditions and 2) cause and effect of fluctuations in biochemical pathways.</td>
</tr>
<tr>
<td>The physical basis of interactions</td>
<td>Interactions occur because of the electrostatic properties of molecules. These properties can involve full, partial, and/or momentary charges.</td>
<td>Once this concept is understood, similarities between different types of interactions become clear. Although interactions are given different names, they are all based on the same electrostatic principles.</td>
<td>A core biochemical principle is that structure governs function. Correct understanding of noncovalent interactions is essential in integrating structure and function.</td>
</tr>
<tr>
<td>Thermodynamics of macromolecular structure formation</td>
<td>Interactions in biological systems almost always take place in aqueous solution. Bulk interactions in an aqueous system have an entropic component. Enthalpic and entropic contributions are responsible for biological structure.</td>
<td>Protein folding, the assembly of lipids into micelles and bilayers, the association of polypeptide subunits to form oligomeric proteins, base pairing of DNA and RNA molecules, and all other biological interactions are driven by a common set of thermodynamic forces. The aqueous environment of the cell plays an active and essential role in biochemical structure formation.</td>
<td>When the entropic and enthalpic forces that drive processes like protein folding and binding are understood, predictions can be made about the conditions under which these events will occur and what effect perturbations, like mutations will have.</td>
</tr>
</tbody>
</table>

Statements shown in Table 4 by comparing student alternate conceptions with expert conceptions generated from a variety of sources, including the faculty workshops, the literature, textbooks, and the authors’ personal knowledge. We next addressed the characteristics of transformative and integrative by asking ourselves what biochemical ideas are unlocked once the concept is understood (transformative) and what previously hidden connections are made visible once the concept is understood (integrative). These ideas (shown in the last two columns on Table 4) were produced through iterative conversations among authors and two project advisory board members. It is important to note that one of
the authors was trained as a physical chemist and was therefore able to contribute a perspective that extended beyond biochemistry. The characteristic of “irreversible” was impossible to define given available data. Table 4 was reviewed in its entirety by two project advisory board members to determine whether the statements were correct and relevant for learning in biochemistry. Modifications were made based on their feedback.

One unanticipated result of this process was that we decided to break the intra- and intermolecular interactions concept into two distinct but related categories: the physical basis of interactions and thermodynamics of macromolecular structure formation. This decision was based on our analysis of student interviews and reflects our vision of how these concepts could be better taught in biochemistry courses. Although students should have a fundamental understanding of Coulomb’s law and intermolecular interactions from general chemistry, responses given during interviews made clear that the students do not have a solid understanding of how these two concepts are related. Student interviewees were unable to make the generalization that all noncovalent interactions are based on charge–charge interactions. They were able to produce apparently memorized lists of interaction types, and some could even describe the polarization of electron clouds. However, when questioned further, they were largely unable to connect the events they had previously described to simple electrostatic interactions. The result of this superficial understanding of intermolecular forces was that they subsequently went on to describe a superficial understanding of these concepts in a biochemical context. For example, many students could list the types of interactions one might find in each level of protein structure, and they could state that hydrophobic groups go to the inside of a protein and hydrophilic to the outside. However, most were unable to identify the atoms involved in interactions in proteins. Furthermore, most students explained the hydrophilic/hydrophobic characteristics by attributing desires to molecules (“they want to be on the inside or outside”). On the basis of these observations, we concluded that classroom strategies aimed at understanding the electrostatic basis of interactions (enthalpic contributions) needs to precede higher-level discussions of structure that rely on enthalpic and entropic contributions.

The more basic discussions should still happen in the context of biological macromolecules.

Finally, a survey was sent to all participants from the interdisciplinary life sciences workshop asking for feedback on the knowledge statements and labels we had given to each set of ideas. Nineteen of 20 participants responded. Although one respondent was concerned that none of the five proposed TCs was narrow enough to fit the definition of a concept, no other participants raised this concern, and no other major concerns were raised. The majority of respondents agreed with or were neutral regarding our choice of names for the five different TCs, and their feedback was used to further refine the names of three of the five TCs to improve clarity and focus.

### DISCUSSION

As has been observed previously with biology TCs, TCs for biochemistry have conceptual connections to each other and to additional biochemical concepts and skills, resulting in a web of ideas (Ross et al., 2010). For example, intermolecular interactions are involved in the vast majority of biochemical events and therefore connect many concepts in biochemistry. Likewise, free-energy changes can be characterized for any biochemical event, and understanding biochemical energetics allows one to make predictions across biochemical contexts. Because of their centrality in biochemistry, the five TCs characterized in Table 4 connect to and underlie the foundational concepts and “big ideas” described in Table 1. This integration emphasizes the central importance of the TCs in learning biochemistry and reaffirms our decision to focus future instructional materials on a small number of knowledge statements related to these concepts.

When a provisional list of TCs was circulated to workshop participants at the conclusion of phase 4, several were concerned by the focus on what they consider to be general chemistry concepts. What has become clear through this undertaking is that we, as biochemistry educators, can no longer act as though students develop deep, correct understandings of key concepts in prerequisite courses. Decades

<table>
<thead>
<tr>
<th>Name</th>
<th>Knowledge statement(s)</th>
<th>Biochemical ideas that are unlocked once this concept is understood</th>
<th>Connections that were invisible before deep understanding of the concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free energy</td>
<td>The tendency toward equilibrium drives biological processes. Differences in free energy drive the chemical transformations underlying biological function. By providing a direct link between a thermodynamically favorable reaction with a thermodynamically unfavorable one, enzymes enable biological systems to drive a normally unfavorable reaction by coupling it to one with a large and favorable free-energy change. Enzymes affect reaction rate, yet do not affect equilibrium position.</td>
<td>Biological systems use favorable processes to drive less-favorable processes, which allows for maintenance of steady state.</td>
<td>Once this concept is understood, the relationship among free energy, equilibrium, and steady state becomes apparent.</td>
</tr>
</tbody>
</table>

---

**Table 4.** Continued
of discipline-based education research on alternate conceptions provide ample evidence that students move through chemistry and biology programs with major gaps in understanding (for examples, see Mulford and Robinson, 2002; Shi et al., 2010; Parker et al., 2012). Unfortunately, these alternate conceptions are often regarded by biochemistry educators as a problem that others need to fix. We assert that this approach is not constructive and that we must develop methods that simultaneously teach new biochemistry concepts, and help students refine and strengthen their understanding of foundational concepts. Furthermore, student interview data suggest that even when students understand concepts in prerequisite courses, they struggle to apply their understanding in new contexts. Separating scientific course work into discipline-specific silos (e.g., chemistry and biology) may act as an additional barrier to student learning of TCs.

Identification of TCs is meaningful only insofar as it is used to inform effective changes in teaching and learning. Therefore, the critical next step is to work with a community of biochemistry educators to produce instructional and assessment materials related to these concepts. For a number of reasons, we are committed to working with a large, nationwide community of biochemistry educators to accomplish this goal. First, there is historical precedent in the TC community to involve a diversity of faculty voices in the process (Cousin, 2009). Furthermore, we have had previous success in developing and disseminating assessment and teaching materials in collaboration with a community of faculty members (Murray et al., 2011; Villafane et al., 2011). Finally, Vision and Change (American Association for the Advancement of Science [AAAS], 2011) and other projects have created national momentum to work collaboratively to reshape teaching and learning in the molecular life sciences. Indeed, the Vision and Change report demands that we engage in regular conversations resulting in a coordinated national effort to change how undergraduates learn in the life sciences.

In our work through the phases of this project, we listened to faculty ideas about the production of teaching materials related to TCs. As a result, we plan to prioritize several guiding principles to support future production of instructional materials. For example, whenever possible, classroom activities should contain components that explicitly develop the visual literacy skills needed for expert interpretation of biochemical representations. These skills were initially considered as a TC in their own right, and the case could readily be made. However, for the purposes of this project, the developing consensus is that it will be more powerful to weave the teaching of visual literacy into as many activities as possible. The literature on visualization in biochemistry will be considered in production of any teaching materials (Schönborn and Anderson, 2006, 2009, 2010; Towns et al., 2012). Another guiding principle is that research skills be included whenever appropriate. These skills could include experimental design and critical interpretation of the literature. There is precedent for including experimental approaches as a cross-cutting concept in course curriculum design. For example, Rowland and colleagues placed experimental approaches at the center of their concept lens diagram to acknowledge that they underlie all knowledge in biochemistry and that developing these skills supports student understanding of the process of science, a key component of Vision and Change (AAAS, 2011; Rowland et al., 2011). Finally, although the relationship between individual and populations of molecules was not determined to be a TC, an effort should be made to make explicit connections between molecular characteristics and observable macroscopic properties.

Curricular changes that support learning of TCs in biochemistry are likely to be multifaceted. Land and colleagues suggest that curricular changes that attempt to focus on TCs should be systematically reviewed to consider three related elements: 1) the sequence in which concepts are introduced and connected, 2) the process by which students recognize and internalize TCs, and 3) the ways in which students and teachers will recognize when TCs have been internalized (Land et al., 2006). They go on to describe nine considerations that are important in curriculum redesign and assessment. Interestingly, their recommendations focus less on what is taught (although discipline-specific concepts and skills are certainly central to any curriculum redesign process) and more on how TCs are taught. They emphasize, among other things, the importance of listening to students, engaging them in the learning process, and teaching students to tolerate uncertainty. Furthermore, they claim the learning of TCs often involves a change in students’ sense of personal identity. Therefore, a curriculum designed around TCs should actively recognize and support attitudinal changes and emotional challenges that could accompany transformational learning. For example, students could be asked to write meta-cognitive self-reflections and receive feedback on these documents from instructors.

In addition to the five concepts described here, the other provisional TCs (Table 3), as well as still-unidentified concepts, are also likely to be instrumental for learning in biochemistry. Therefore, although we are confident that the five targeted areas are indeed TCs for biochemistry, they by no means represent an exhaustive list and may not even be the five most important TCs for biochemistry. Rich opportunities remain to further refine the provisional list using the process described in this paper. A long-term goal would be to reimagine how content is organized in undergraduate biochemistry courses, such that course structure aligns with deep learning of TCs. A possible model comes from Rowland and coworkers, who rearranged their course content using concept lenses, a model that arose from the “big ideas” in the molecular life sciences (Rowland et al., 2011). This model discards traditional organizational schemes for biochemistry courses, which are often based on textbook organization and content choices, and replaces them with a student-centered, concept-based structure that seeks to build deep conceptual understanding and habits of mind used by professional biochemists.

CONCLUSIONS

We have undertaken a nationwide effort to identify TCs for biochemistry with the goal of improving student learning in undergraduate biochemistry courses. Five TCs—steady state, biochemical pathway dynamics and regulation, the physical basis of interactions, thermodynamics of macromolecular structure formation, and free energy—will be the target of future work to produce instructional and assessment materials with the aim of improving learning in undergraduate biochemistry courses. The targeted TCs are integral to deep
understanding of biochemistry but also relate strongly to foundational concepts from general chemistry and biology. Our inability to fully disaggregate learning in biochemistry from foundational concepts in prerequisite courses emphasizes the importance of holistic curriculum design both within and among disciplines.

ACKNOWLEDGMENTS
The work described in this article was supported by NSF DUE-1224868. 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. Thanks to our advisory board members, Peter Kennelly, Daniel Schwartz, Robin Wright, and Adele Wolfson. Thanks to all of the faculty members who participated in this project. A special thanks to faculty members who attended 2- or 4-d workshops: Kevin Aheron, Karen Anderson, Rodney Austin, Adam Cassano, Colleen Conway, Carol Deckman, Martina Edener, Matthew Fisher, Katherine Frato, January Haile, Bruce Heyen, Margaret Kanipes, Jennifer Knight, Lisa Kroutil, Anne Kruchten, Paula Lemons, Kimberly Linenerber, Sunil Malapati, Jenny McFarland, Tracey Murray, Philip Ortiz, Jeffrey Owens, David Parkin, Mandy Schivell, Janet Schottel, Duane Sears, Tricia Shepherd, Daniel Smith, Kimberly Tanner, Ann Taylor, Heather Tienson, Sachel Villañafe, Mark Werth, Adele Wolfson, and several participants who wished to remain anonymous. We appreciate the help of Trey Morgan, a student who helped with data processing. Finally, we thank all of the students who participated in focus groups and shared their understanding and insights with us.

REFERENCES
Clouder L (2005). Caring as a “threshold concept” transforming students in higher education into health(care) professionals. Teaching High Educ 10, 505–517.


White HB, Benore MA, Sumter TF, Caldwell BD, Bell E (2013). What skills should students of undergraduate biochemistry and molecular biology programs have upon graduation? Biochem Mol Biol Educ 41, 297–301.


Redefining Authentic Research Experiences in Introductory Biology Laboratories and Barriers to Their Implementation

Rachelle M. Spell,* Judith A. Guinan,† Kristen R. Miller,‡ and Christopher W. Beck*§

*Emory University, Atlanta, GA 30322; †Radford University, Radford, VA 24142; ‡University of Georgia, Athens, GA 30602; §Flinders University, Adelaide, SA 5001, Australia

Submitted October 13, 2013; Revised December 11, 2013; Accepted December 14, 2013

Monitoring Editor: Nancy Pelaez

Incorporating authentic research experiences in introductory biology laboratory classes would greatly expand the number of students exposed to the excitement of discovery and the rigor of the scientific process. However, the essential components of an authentic research experience and the barriers to their implementation in laboratory classes are poorly defined. To guide future reform efforts in this area, we conducted a national survey of biology faculty members to determine 1) their definitions of authentic research experiences in laboratory classes, 2) the extent of authentic research experiences currently experienced in their laboratory classes, and 3) the barriers that prevent incorporation of authentic research experiences into these classes. Strikingly, the definitions of authentic research experiences differ among faculty members and tend to emphasize either the scientific process or the discovery of previously unknown data. The low level of authentic research experiences in introductory biology labs suggests that more development and support is needed to increase undergraduate exposure to research experiences. Faculty members did not cite several barriers commonly assumed to impair pedagogical reform; however, their responses suggest that expanded support for development of research experiences in laboratory classes could address the most common barrier.

INTRODUCTION

Over the past 15 yr, numerous reports have called on science faculty to bridge research and teaching in order to attract more undergraduates to science, technology, engineering, and mathematics (STEM) fields (Boyer Commission on Educating Undergraduates in the Research University, 1998; Project Kaleidoscope, 2002; National Research Council [NRC], 2003, 2005). More recently, the American Association for the Advancement of Science (AAAS) Vision and Change report urged the integration of research experiences into biology education for all students (AAAS, 2011). The goal of these reforms is authentic science instruction that reflects better what scientists do and how scientists think. These directives are based on many studies showing the positive impact of traditional undergraduate research experiences, alternatively called mentored, independent, or collaborative research, in which a student works in a laboratory or in the field on a project with some faculty oversight (reviewed by Seymour et al., 2004; Lopatto, 2006; Russell et al., 2007; Crowe and Brakke, 2008; Laursen et al., 2010). The number of students who can have such a traditional research experience is limited, however, because of factors such as lack of funding (President’s Council of Advisors on Science and Technology, 2012). In contrast, all science students take laboratory classes. Unfortunately, the traditional laboratory experience in lab classes, with “cookbook” experiments in which students follow a predetermined methodology and then prepare an artificial scientific report, is not effective for teaching scientific content or the process of science (Myers and Burgess, 2003; Luckie et al., 2004; Rissing and Cogan, 2009). Such canned activities can foster an unrealistic view of science as prescriptive, unambiguous, and irrelevant and therefore do not reflect “authentic” science (Bencze and Hodson, 1999; Rahm et al., 2003; NRC, 2005).
Because research-oriented laboratory courses can provide many of the same advantages as traditional mentored-research opportunities (Lopatto, 2009) and can reach more students, recent efforts have been made to introduce authentic research experiences in laboratory courses. Attempts to incorporate the process of science into the lab exercises have resulted in several different lab course models that vary in level of student independence and in the types of research questions addressed (examples are reviewed in Graham et al., 2013). These different course-based research projects focus on different aspects of authentic research. If support for research experiences is to be expanded to laboratory instruction, it is important to define the essential components of an authentic research experience in a laboratory course.

The essential components of mentored research include reading scientific literature, designing aspects of the research project, working toward significant findings, and oral and written presentation of the results (Seago, 1992; Lopatto, 2003). Several publications have attempted to define the level of research experiences in laboratory class, using terms like inquiry (e.g., guided inquiry, open-ended inquiry, teacher-collaborative inquiry, and authentic inquiry), authentic, or research-based (D’Avanzo, 1996; Olson and Loucks-Horsley, 2000; Rahm et al., 2003; Buck et al., 2008; Weaver et al., 2008; Bencze and Hodson, 2009). However, the use of any one definition may not reflect the vision of biology faculty about incorporating authentic research experiences in the classroom. Education reform toward authentic science instruction includes definitions of authentic science as a process (involving critical thinking, hypothesis generation, experimental design, data analysis, and scientific communication) or as a product (novel results, new discoveries, and publishable data). This lack of a clear definition or adherence to a definition not universally held could impede acceptance, implementation, and funding for course-based research experiences.

Other barriers to pedagogical reform could limit the expansion of research experiences into laboratory courses. Inadequate training, time, and incentive are among the most commonly cited barriers for faculty change to reform teaching (for a review of the literature, see Brownell and Tanner, 2012). Additional potential barriers include struggles with professional identity (Brownell and Tanner, 2012); logistical issues, such as adequate funding and persons to implement teaching reform (Kloser et al., 2011); and faculty attitude that traditional teaching formats are sufficient, because the students who experience such formats do graduate and get jobs (Benvenuto, 2002).

Efforts to broaden exposure to science inquiry through authentic research experiences in laboratory courses could be hampered by assumptions and lack of information about the essential components of authentic research experiences in these courses and the barriers to their implementation. Henderson et al. (2011) note that, to bring about change in pedagogical approaches, those changes must align with the pedagogical values of faculty involved. To support faculty members’ efforts to answer calls for research for all biology students, it is important to gauge their views of the critical components of experiences that mimic authentic research, determine the extent of current implementation of authentic research in courses, and identify the barriers that prevent the expansion of those experiences.

Previous surveys of laboratory courses have explored the types of experiences offered in introductory biology laboratory classes. A survey of members of the Ecological Society of America at 131 institutions found a very low level of open-ended investigations in 1996 (Brewer, 1998). The most recent report of the status of introductory biology laboratories was in 2005, with a survey of 65 institutions, asking faculty to report whether lab experiences incorporated inquiry-based or open-ended investigations (Sundberg and Armstrong, 1993; Sundberg et al., 2005). Sundberg and colleagues found a high level of inquiry-based experiences in all institution types, but open-ended investigations were less common, especially in research and comprehensive institutions. The information from these surveys, now outdated, provided no information about the degree to which these research approaches are used in a laboratory course nor the faculty members’ views of the critical components of research in laboratory classes.

Research Experiences in Introductory Laboratories (REIL)—Biology is a network of faculty interested in exploring questions related to authentic research in introductory biology laboratory courses. In particular, we want to expand professional development opportunities for faculty to create authentic research experiences in laboratory classes. To inform efforts to provide authentic research experiences for all students through laboratory courses, we surveyed instructors nationally to determine their views on the essential components of authentic research experiences in introductory biology laboratory courses, the level to which these experiences are incorporated in these courses, and the perceived barriers to implementing them.

METHODS

Survey Development and Implementation

We developed and carried out a national survey on authentic research experiences in introductory biology laboratory courses (see the Supplemental Material). The survey was developed at a meeting of 12 faculty members involved in the REIL—Biology Research Coordination Networks in Undergraduate Biology Education (RCN-UBE) incubator project (http://rcn.ableweb.org), along with core participants in the Introductory Biology Project (http://ibp.ou.edu). The survey asked faculty to define the essential components of an authentic research experience in a laboratory course. Based on their definitions, faculty members were asked to provide demographic information about the introductory biology laboratory courses they taught at their institutions, including the percentage of those courses devoted to authentic research, and to rank the importance of particular barriers to the implementation of authentic research. The survey was widely disseminated through email lists of the biology education community, including the education sections of disciplinary societies, the American Institute of Biological Science, the Association for Biology Laboratory Education, and the Council on Undergraduate Research (CUR).

Survey Response Coding and Analysis

Coding Authentic Research Definitions. Definitions of authentic research were coded in a two-step process. First, NVivo software (QSR International, Burlington, MA) was
used to create a list of the most frequently provided words (and variations thereof) in the definitions of authentic research. These terms and their variations were used to create broad categories of descriptive expressions common to the majority of definitions. The categories included the words: “experiment,” “data,” “question,” “research,” “design,” “hypothesis,” “presentation,” and “results.” We then used the process of thematic analysis to refine the categories and capture specific themes (and their frequencies) that occurred within them. This process allowed categories to be considered within the context of entire responses rather than just in light of frequency of individual terms. For example, the word “experiment” was the most common term found across all definitions. However, deeper examination clarified that, in most cases, survey respondents were referring to students creating their own experimental designs to answer a question.

To begin the analysis, one of us (C.W.B.) reviewed all definitions and categories and provided a preliminary coding scheme of six themes within the definitions. These included novel questions (i.e., unknown results), student-generated questions, hypothesis formation, experimental design, data analysis, and presentation or publication. Two of us (R.M.S. and K.R.M.) then coded a random subset of definitions (n = 40) using the themes. We used Krippendorff’s alpha to compare the interrater reliability of applied codes. Themes that resulted in low interrater reliability were discussed until mutual agreement was reached on understanding what a theme represented and how it should be coded. Finally, three of us (C.W.B., R.M.S., and K.R.M.) coded the remaining definitions. For all themes, the interrater reliability was greater than 0.70. Therefore, for definitions not coded the same by the three of us, we used the coding of the majority. During the coding, a seventh theme, data collection, emerged and was coded for all definitions. Eighty-eight percent of the definitions included at least one of the seven themes. Many of the definitions that did not fit the defined themes were caused by the respondents’ failure to directly address the question.

Identification of Additional Barriers. An open-ended portion of the survey allowed respondents to provide additional barriers to implementation of authentic research experiences. Eighty-six responses were recorded; some responses included multiple additional barriers. Seventy-eight instances were examples or elaborations of barriers listed in the survey. However, 22 new, distinct barrier categories also emerged from the data, and frequencies of these barriers were recorded.

Statistical Analyses

Definitions of Authentic Research. We used a Spearman correlation test to determine the relationship between the themes that were suggested by respondents’ definitions of authentic research. To determine whether the presence or absence of particular themes differed significantly among institution types, we performed a Kruskal-Wallis rank-sum test. A Kruskal-Wallis rank-sum test was used to determine whether the number of themes selected varied significantly based on institution types.

Factors Influencing the Percent Authentic Research Incorporated into Laboratory Exercises. Because the variable percent research was nonnormally distributed and attempts to transform it were ineffective, we analyzed the factors influencing the amount of authentic research incorporated into laboratories using nonparametric methods. We used a Spearman’s correlation to examine the relationship between percent research and the number of course sections for each course reported in the surveys. To compare the percent research in courses (averaged across all courses reported by each respondent) taught by individuals who included a particular theme with those who did not include that theme, we used Wilcoxon rank-sum tests. We determined whether the percent research varied according to the number of themes suggested by respondents’ definition of authentic research using a Kruskal-Wallis rank-sum test. To determine the relationship between institution or course type and percent research, we averaged responses across course type for each respondent and binned them into categories as follows: 0%, 0–25%, 26–50%, 51–75%, and 76–100%. We then determined the relationship between the binned categories and institution and course type using Pearson χ² analyses.

Analysis of Reported Barriers to Including Authentic Research Experiences. We used Kruskal-Wallis rank-sum tests to examine the influence of institution type (e.g., research university, 2-yr college, etc.) on the distribution of responses for each of the barriers. Differences in perceptions of barriers between faculty from minority and non-minority serving institutions were tested using Wilcoxon rank-sum tests. Similarly, we compared the perceptions of barriers of faculty from private and public institutions using Wilcoxon rank-sum tests. Finally, we used Wilcoxon rank-sum tests to determine whether there were differences in perceived barriers depending on whether particular themes were included in faculty definitions of authentic research.

All analyses were performed using JMP, version 9.0.0 statistical software (SAS Institute, Cary, NC).

RESULTS

Sample

Our national survey elicited 279 responses across institution types (2-yr college: 17%; liberal arts college: 41%; comprehensive university: 13%; research university: 20%; not stated: 9%), equally split between public and private institutions. Fourteen percent of responses were from minority-serving institutions. Respondents provided data on 534 courses, the majority of which were courses for majors (Figure 1). The vast majority of respondents were leaders in these courses (76% instructors and 15% directors). The frequency of responses for courses serving majors, nonmajors, and mixed (majors and nonmajors) did not differ among institution types (χ² = 6.38, df = 6, p = 0.39). However, courses for prehealth students were more common at 2-yr colleges in our sample (Figure 1).

Definitions of Authentic Research

A word-frequency search of open-ended responses to our question about the essential components of an authentic research experience in a laboratory course identified seven common themes (in order of decreasing frequency): 1) experimental design (n = 145), 2) data collection (n = 136), 3) data analysis (n = 131), 4) presentation or publication (n = 103), 5) hypothesis formation (n = 89), 6) student-generated

Vol. 13, Highlights of 2014 91
questions \((n = 83)\), and 7) novel questions \((n = 62)\). Fewer than 15% of definitions failed to fit at least one of these themes, often because of confusion over the question being asked. Most responses included more than one theme (Figure 2). The number of themes (Kruskal-Wallis test, \(\chi^2 = 7.72, df = 3, p = 0.052\)) and the presence or absence of particular themes (chi-square goodness-of-fit tests, \(p > 0.05\)) were not significantly affected by institution type.

Two conceptions of authentic research in laboratory courses emerged. The more common conception emphasized the process of science, often incorporating multiple themes except the novel questions theme (Figures 2 and 3). The less common conception emphasized novel research questions to the exclusion of the other themes (Figure 3). In other words, when a respondent included only a single theme, it was most often the novel questions theme (Figure 2). In most cases, the themes were significantly positively correlated (Table 1), but the presence of the novel questions theme in a response was either significantly negatively correlated or uncorrelated with the inclusion of all other themes (Table 1), supporting the idea of two different conceptions of authentic research in laboratory courses (Figure 3).

**Level of Authentic Research**

Students in introductory biology laboratory courses spend on average a third of their time on authentic research activities; however, the amount of authentic research incorporated into individual courses is variable, ranging from 0 to 100% (Figure 4A). Twenty-three percent (23%) of courses have no
research, and 56% devote less than a quarter of course time to research. The amount of research in a course is not affected by the number of themes included in a respondent’s definition of authentic research (Kruskal-Wallis test, \( \chi^2 = 5.28, df = 6, p = 0.44 \)) or which themes were included (Wilcoxon rank-sum test, \( p > 0.15 \) for all themes). However, research is much less common in nonmajors courses (\( \chi^2 = 26.93, df = 1, p < 0.001 \)) and at 2-yr colleges (\( \chi^2 = 9.87, df = 1, p = 0.002; \) Figure 4B). The percentage of research in a course also decreases as the number of sections of a course increases (Spearman rank correlation, \( r = -0.167, p = 0.0002 \)).

### Barriers to Implementing Authentic Research

A summary of the most and least important barriers based on the most frequent faculty response (mode) is given in Table 2 (see Supplemental Table S1 for overall summary statistics and Tables S2–S4 for summary statistics by institution type). Across all institution types, lack of time to develop new laboratory research experiences was reported as the biggest barrier to implementation of authentic research in laboratory courses for most faculty (mode of 4 out of 4 on a Likert scale where 1 = “not a barrier,” 2 = “minor barrier,” 3 = “somewhat a barrier,” and 4 = “major barrier”; Tables 2 and S1). Furthermore, additional barriers were considered to be major barriers by faculty at specific institution types (mode = 4; Tables 2 and S2–S4). In some cases, a potential barrier was, in fact, not a barrier (mode = 1; Table S1) when we examined the data across all faculty but was considered a major barrier (mode = 4) for a subset of faculty (e.g., lack of administrator support and instructor resistance at minority-serving institutions; Table S3).

When we compared the importance of potential barriers between institution types statistically, we found the reports of several barriers differed significantly (Tables S2–S4). For example, faculty at 2-yr colleges indicated that lack of time for faculty to develop new research activities and lack of student preparation were significant barriers, whereas faculty at research universities identified large class sizes and large number of sections to be significant barriers (Table S2). Faculty at minority-serving institutions identified lack of administrator support, lack of facilities, and lack of time to develop new research experiences as significant barriers (Table S3). Cost, lack of equipment, number of sections, and lack of student preparation are perceived to be significantly greater barriers by faculty at public institutions in contrast to faculty at private institutions (Table S4). The barriers due to number of sections and lack of student preparation exhibit the most noticeable differences in modal response between faculty at public and

| Table 1. Spearman rank correlations between different themes in the definitions of authentic research \(^a\) |
|-----------------|----------------|----------------|----------------|----------------|----------------|
|                 | Novel questions | Student-generated questions | Hypothesis formation | Presentation or publication | Data analysis |
| Novel questions | 1               | -0.18            | -0.20           | 0.04            | -0.11          | -0.21          | -0.16          |
| Student-generated questions | -0.18 | 1 | -0.01 | 0.04 | 0.11 | 0.13 | 0.25 |
| Hypothesis formation | -0.20 | -0.01 | 1 | 0.15 | 0.27 | 0.21 | 0.06 |
| Presentation or publication | 0.04 | 0.04 | 0.15 | 1 | 0.37 | 0.43 | 0.20 |
| Data analysis | -0.11 | 0.11 | 0.27 | 0.37 | 1 | 0.50 | 0.27 |
| Data collection | -0.21 | 0.13 | 0.21 | 0.43 | 0.50 | 1 | 0.22 |
| Experimental design | -0.16 | 0.25 | 0.06 | 0.20 | 0.27 | 0.22 | 1 |

\(^a\)Bolded values are significantly correlated \((p < 0.05)\).
private institutions (Table S4). Additional barriers listed in free responses fell into multiple, small categories \((n = 22)\), including time for implementation, logistics, student attitudes, connection with lecture, creativity, and curricular resources (Table S5), which suggests a strong contextual influence on implementation of authentic research experiences. Overall, the primary barriers to implementing authentic research (Table S1) did not differ depending on how a respondent defined authentic research, except that those faculty members who included the theme of experimental design considered lack of student preparation to be more of a barrier \((\chi^2 = 4.35, df = 1, p = 0.037)\).

Some of the potential barriers included in the survey were not considered to be barriers by faculty respondents. Instructor resistance, lack of administrator support, lack of facilities, effects on student evaluation of instructors, loss of content coverage and breadth, lack of ways to effectively assess student were all most often perceived to be not a barrier (mode = 1; Tables 2 and S1). In particular, more than half of the respondents considered lack of administrator support and effects on student evaluation of instructors not to be a barrier (median; Table S1). However, it is important to note that the degree to which some of these potential barriers were considered to be (un)important varied among institution types (Table S2), between minority and nonminority institutions (Table S3), and between public and private institutions (Table S4).

**DISCUSSION**

This study examined the definitions of authentic research in the context of introductory biology laboratory courses, the level of research experiences in these courses, and the barriers to implementing authentic research according to faculty members who teach these courses. Surprisingly, our survey found different understandings of the essential components of authentic research and highlighted errors in conventional wisdom about barriers to implementing research experiences.
in laboratory courses. The survey sampled a large number of faculty members representing many courses in all institutions types, but, of course, the results could be biased by differences between those who responded versus those who did not. It is possible, for example, that faculty members who offer research opportunities in laboratory courses were more likely to respond, inflating the percent of laboratory instruction that involves research experiences reported in our study.

Also, conceptions of authentic research and barriers to its implementation likely vary from individual to individual and institution to institution. Finally, characteristics of authentic research experiences may also be different in upper-level courses. In short, there is no one “best” definition of authentic research and no single barrier to overcome. Thus, this study emphasizes the need to discuss definitions and measure barriers before directing education reform. If we do not, we run the risk of wasting efforts and ignoring real barriers. In a time of federal reevaluation of best practices for supporting STEM education reform (Mervis, 2013), such analyses are even more critical.

The results of our survey suggest two conceptions of the essential components of authentic research in introductory biology laboratory courses (Figure 3). The predominant conception relates to the process of science. In contrast, the less frequent conception emphasizes exclusively novel research questions. Clearly, the essential components of research experiences in laboratory courses that emerged from our survey overlap with those defined by Lopatto (2003) for mentored research experiences. However, given the emphasis on science process over novel results in most responses, all faculty do not perceive authentic research in a laboratory course as simply migrating mentored research into a course context and scaling it up. The intention of this study is not to choose which conception is correct but rather to measure the prevalence of both views.

These two conceptions likely arise in part because of the difference in their primary goals. Whereas the main objective of mentored research is to produce science (i.e., publishable, novel results), the main objective of research in laboratory courses is to produce scientists (i.e., students who understand the process of science). This dichotomy has been described as a conflict between “products” versus “process” (Rahm et al., 2003) or “science-as-a-noun” versus “science-as-a-verb” (Apedoe et al., 2006). Both goals still reflect authentic science instruction, in that they reflect what scientists do in their own research laboratories. Although these goals are distinct, they do not have to be exclusive; a well-mentored research lab should strive to develop future scientists while generating new data. Efforts to increase authentic research in laboratory courses need to recognize that definitions of authentic research are context dependent. Faculty who implement research courses under different goals should be sensitive to the limitations of either conception. Thus, faculty members who develop course-based research that emphasizes student-generated questions and experimental design but may not generate publishable data could provide extra discussion of the rigor of the scientific process and publication during their courses. Likewise, faculty members who oversee research activities that generate publishable data, often addressing a faculty-defined hypothesis using established protocols, could intentionally include exercises in student experimental design. Though some models of authentic research in courses that include both conceptions exist, they are rare, and it would be unrealistic to require both in all models of course-based research activities.

Different successful models of authentic research experiences in laboratory courses demonstrate the diversity of approaches. Previous efforts have developed specific model systems for research-based laboratory courses (e.g., Research Link 2000, www.cur.org/projects_and_services/research_link2000). In addition, model projects implementing research experiences in college laboratory courses are being used across multiple institutions (e.g., the Genome Consortium for Active Teaching [Campbell et al., 2006, 2007], Integrated Microbial Genomes Annotation Collaboration Toolkit [Ditty et al., 2010], Howard Hughes Medical Institute [HHMI] Phage Hunters [Hanauer et al., 2006], Genomics Education Partnership [Shaffer et al., 2010], Ecological Research as Education Network [Bowne et al., 2011]). Several examples of research laboratory courses explore topics that derive from the instructors’ research programs (e.g., Bell, 2011; Gardner et al., 2011; Gasper et al., 2012). Many of these projects focus predominantly on the conception of authentic research related to novel results, in which the research question or the methods are defined by the instructor. The Course-Based Undergraduate Research Experience Network is working to bring together many of these models. On the other hand, models of laboratory activities incorporating student-generated questions and experimental design are commonly described in the literature as open-ended inquiry. In many cases, these exercises do not generate publishable data.

Some examples of authentic research in courses show ways in which the two different conceptions of authentic research

### Table 2. Summary of barriers to implementation of authentic research experiences in introductory biology laboratory classes, based on Likert-scale modes

<table>
<thead>
<tr>
<th>Overall</th>
<th>Most important barrier</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Lack of time for faculty to develop new research experiences</td>
</tr>
<tr>
<td></td>
<td>• Instructor resistance</td>
</tr>
<tr>
<td></td>
<td>• Lack of administrator support</td>
</tr>
<tr>
<td></td>
<td>• Lack of facilities</td>
</tr>
<tr>
<td></td>
<td>• Effects on student evaluation of instructors</td>
</tr>
<tr>
<td></td>
<td>• Loss of content coverage and breadth</td>
</tr>
<tr>
<td></td>
<td>• Lack of ways to effectively assess students</td>
</tr>
<tr>
<td>Not barriers</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Lack of student preparation</td>
</tr>
<tr>
<td></td>
<td>• Number of sections</td>
</tr>
<tr>
<td></td>
<td>• Cost</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Lack of student preparation</td>
</tr>
<tr>
<td></td>
<td>• Instructor resistance</td>
</tr>
<tr>
<td></td>
<td>• Lack of administrator support</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Number of sections</td>
</tr>
<tr>
<td></td>
<td>• Class size</td>
</tr>
<tr>
<td></td>
<td>• Class size</td>
</tr>
<tr>
<td></td>
<td>• Lack of student preparation</td>
</tr>
<tr>
<td></td>
<td>• Lack of student preparation</td>
</tr>
<tr>
<td></td>
<td>• Lack of student preparation</td>
</tr>
<tr>
<td></td>
<td>• Lack of student preparation</td>
</tr>
<tr>
<td></td>
<td>• Lack of student preparation</td>
</tr>
<tr>
<td></td>
<td>• Lack of student preparation</td>
</tr>
<tr>
<td></td>
<td>• Lack of student preparation</td>
</tr>
<tr>
<td></td>
<td>• Lack of student preparation</td>
</tr>
<tr>
<td></td>
<td>• Lack of student preparation</td>
</tr>
<tr>
<td></td>
<td>• Lack of student preparation</td>
</tr>
<tr>
<td></td>
<td>• Lack of student preparation</td>
</tr>
<tr>
<td></td>
<td>• Lack of student preparation</td>
</tr>
</tbody>
</table>

The results of our survey suggest two conceptions of the essential components of authentic research in introductory biology laboratory courses (Figure 3). The predominant conception relates to the process of science. In contrast, the less frequent conception emphasizes exclusively novel research questions. Clearly, the essential components of research experiences in laboratory courses that emerged from our survey overlap with those defined by Lopatto (2003) for mentored research experiences. However, given the emphasis on science process over novel results in most responses, all faculty do not perceive authentic research in a laboratory course as simply migrating mentored research into a course context and scaling it up. The intention of this study is not to choose which conception is correct but rather to measure the prevalence of both views.

These two conceptions likely arise in part because of the difference in their primary goals. Whereas the main objective of mentored research is to produce science (i.e., publishable, novel results), the main objective of research in laboratory courses is to produce scientists (i.e., students who understand the process of science). This dichotomy has been described as a conflict between “products” versus “process” (Rahm et al., 2003) or “science-as-a-noun” versus “science-as-a-verb” (Apedoe et al., 2006). Both goals still reflect authentic science instruction, in that they reflect what scientists do in their own research laboratories. Although these goals are distinct, they do not have to be exclusive; a well-mentored research lab should strive to develop future scientists while generating new data. Efforts to increase authentic research in laboratory courses need to recognize that definitions of authentic research are context dependent. Faculty who implement research courses under different goals should be sensitive to the limitations of either conception. Thus, faculty members who develop course-based research that emphasizes student-generated questions and experimental design but may not generate publishable data could provide extra discussion of the rigor of the scientific process and publication during their courses. Likewise, faculty members who oversee research activities that generate publishable data, often addressing a faculty-defined hypothesis using established protocols, could intentionally include exercises in student experimental design. Though some models of authentic research in courses that include both conceptions exist, they are rare, and it would be unrealistic to require both in all models of course-based research activities.

Different successful models of authentic research experiences in laboratory courses demonstrate the diversity of approaches. Previous efforts have developed specific model systems for research-based laboratory courses (e.g., Research Link 2000, www.cur.org/projects_and_services/research_link2000). In addition, model projects implementing research experiences in college laboratory courses are being used across multiple institutions (e.g., the Genome Consortium for Active Teaching [Campbell et al., 2006, 2007], Integrated Microbial Genomes Annotation Collaboration Toolkit [Ditty et al., 2010], Howard Hughes Medical Institute [HHMI] Phage Hunters [Hanauer et al., 2006], Genomics Education Partnership [Shaffer et al., 2010], Ecological Research as Education Network [Bowne et al., 2011]). Several examples of research laboratory courses explore topics that derive from the instructors’ research programs (e.g., Bell, 2011; Gardner et al., 2011; Gasper et al., 2012). Many of these projects focus predominantly on the conception of authentic research related to novel results, in which the research question or the methods are defined by the instructor. The Course-Based Undergraduate Research Experience Network is working to bring together many of these models. On the other hand, models of laboratory activities incorporating student-generated questions and experimental design are commonly described in the literature as open-ended inquiry. In many cases, these exercises do not generate publishable data.

Some examples of authentic research in courses show ways in which the two different conceptions of authentic research
Recent research suggests that the professional identity of faculty as researchers rather than teachers serves as a barrier to change in pedagogy. Professional identity was not included as one of the potential barriers in our survey (see the Supplemental Material). However, it was not listed as an additional barrier by any of our survey respondents (Table S5). Yet integrating a faculty member’s research program into a laboratory course would be an ideal way in which to overcome this potential barrier (e.g., Chen et al., 2005; Kloser et al., 2011; Fukami, 2013; Miller et al., 2013).

On the basis of the results of our survey, we propose several ways in which to address the challenge of the Vision and Change report to provide research experiences in biology for all students. First, to expand efforts by researchers to introduce research into the curriculum, research grant supplements should be offered to scientists (similar to Research Experiences for Undergraduates supplements) to partner with science faculty with education specialties (SFES; Bush et al., 2013), postdoctoral fellows, or graduate students to develop research-based curricula for introductory laboratories for use at their institution or as modules for broader dissemination. Second, to expand the opportunity for education specialists to develop such exercises, funded fellowships should be made available for SFES to work with researchers. Last, to address the most common barrier to implementing authentic research experiences, faculty need time for professional and curriculum development, for example workshops funded by RCN-UBE, Transforming Undergraduate Education in Science, Technology, Engineering and Mathematics, and HHMI. In all cases, faculty members participating in these programs need to begin by defining the essential components of an authentic research experience for their contexts and identifying the barriers that need to be addressed.

ACKNOWLEDGMENTS

We thank members of the REIL-Biology network for their contributions in crafting the survey and providing feedback on a draft of the manuscript. We also thank Arri Eisen for his comments. All data reported in the paper are presented in the Supplemental Material. All authors contributed to writing the survey and the final manuscript. J.A.G. conducted the statistical analyses. R.M.S., K.R.M., and C.W.B. coded the responses to the question on definitions of authentic research. This research was supported by grant DBI-1061798 from the National Science Foundation (NSF) to Emory University. 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


At the college level, the effectiveness of active-learning interventions is typically measured at the broadest scales: the achievement or retention of all students in a course. Coarse-grained measures like these cannot inform instructors about an intervention’s relative effectiveness for the different student populations in their classrooms or about the proximate factors responsible for the observed changes in student achievement. In this study, we disaggregate student data by racial/ethnic groups and first-generation status to identify whether a particular intervention—increased course structure—works better for particular populations of students. We also explore possible factors that may mediate the observed changes in student achievement. We found that a “moderate-structure” intervention increased course performance for all student populations, but worked disproportionately well for black students—halving the black–white achievement gap—and first-generation students—closing the achievement gap with continuing-generation students. We also found that students consistently reported completing the assigned readings more frequently, spending more time studying for class, and feeling an increased sense of community in the moderate-structure course. These changes imply that increased course structure improves student achievement at least partially through increasing student use of distributed learning and creating a more interdependent classroom community.

INTRODUCTION

Studies across the many disciplines in science, technology, engineering, and mathematics (STEM) at the college level have shown that active learning is a more effective classroom strategy than lecture alone (reviewed in Freeman et al., 2014). Given this extensive evidence, a recent synthesis of discipline-based education research (DBER; Singer et al., 2012) suggests that it is time to move beyond simply asking whether or not active learning works to more focused questions, including how and for whom these classroom interventions work. This type of research is being referred to as second-generation education research (Eddy et al., 2013; Freeman et al., 2014) and will help refine and optimize active-learning interventions by identifying the critical elements that make an intervention effective. Identifying these elements is crucial for successful transfer of classroom strategies between instructors and institutions (Borrego et al., 2013).

Using these DBER recommendations as a guide, we have replicated a course intervention (increased course structure; Freeman et al., 2011) that has been demonstrated to increase student achievement at an R1 university and explored its effectiveness when transferred to a different university with a different instructor and student population. Specifically, we expanded on the original intervention studies by exploring 1) how different student subpopulations respond to the treatment in terms of achievement and 2) course-related behaviors and perceptions. These two forms of assessment will help us both elucidate how this intervention achieves the observed increases in student achievement and identify the elements critical for the intervention’s success.
Are Active-Learning Interventions Transferable?

The transferability of active-learning interventions into novel educational contexts is critical to the successful spread of active learning across universities (National Science Foundation, 2013). Unfortunately, transferability of an intervention across contexts cannot be assumed, as there is some evidence that the success of classroom interventions depends on the student populations in the classroom (Brownell et al., 2013), instructor classroom management style (Borrego et al., 2013), and the topics being taught (Andrews et al., 2011). Thus, interventions that work with one instructor at one institution in one class may not necessarily transfer into novel contexts. Yet the majority of published active-learning interventions at the college level have been tested with at best one or two instructors who are usually at the same institution.

We test the transferability of the increased course structure intervention (Freeman et al., 2011), which was effective at a Pacific Northwest R1 university with a predominately white and Asian student body, in a Southern R1 university with a different instructor (who had no contact with the original authors) and a more diverse student body. Additionally, the original study was an introductory biology course for aspiring majors, while the current implementation included mostly nonmajors in a mixed-majors general education course. Thus, in this study, we test the transferability of the increased course structure intervention across three contexts: 1) different instructors, 2) different student body, and 3) different courses (majors vs. nonmajors).

Do Course Interventions Differentially Impact Achievement in Some Student Subpopulations?

There is emerging evidence that classroom interventions could have different impacts on students from different cultural contexts. For example, Asian-American students learn less when they are told to talk through problems out loud compared with when they think through them silently. White students, on the other hand, performed just as well, and in some cases better, when allowed to talk through problems (Kim, 2002, 2008). This finding has implications for a differential impact of peer instruction on Asian students relative to their white classmates. In addition to different cultural norms for learning, students from different subpopulations bring different value sets into the classroom that can influence how they learn in different classroom environments. For example, one study found that when a setting is perceived as interdependent (rather than independent) first-generation students perform better, but continuing-generation students do not differ (Stephens et al., 2012). Positive interpersonal feelings also increased the performance of Mexicans but not European Americans on a learning task (Savani et al., 2013). Thus, the classroom environment itself could have differential impacts on different students. Findings like these begin to call into question whether “one-size-fits-all” classrooms interventions are possible and encourage researchers to disaggregate student response data by subpopulations (Singer et al., 2012).

Up until now, the majority of college-level program evaluations that have disaggregated student groups have done so broadly based on their historical presence in science (underrepresented minority [URM] vs. majority students). Also, most of these studies have explored the impact of supplemental instruction outside an actual science course on student achievement (reviewed in Tsui, 2007; Fox et al., 2009). Only a few STEM course-based curricular interventions have disaggregated student performance (physics: Etkina et al., 1999; Hitt et al., 2013; math: Hooker, 2010; physical science: Poelzer and Zeng, 2008). In biology, two course-based active-learning interventions have been shown to reduce achievement gaps between historically underrepresented students and majority students. Preszler (2009) replaced a traditional course (3 h of lecture each week) with a reformed course that combined 2 h of lecture with 1 h of peer-led workshop. This change in class format increased the grades of all participating students, and the performance of URM students and females increased disproportionately. The second intervention was the increased course structure intervention (Haak et al., 2011). This intervention decreased the achievement gap between students in the Educational Opportunities Program (students from educational or economically disadvantaged backgrounds) and those not in the program by 45% (Haak et al., 2011).

Studies that cluster students into two categories (URM vs. majority) assume that students within these clusters respond in the same way to classroom interventions. Yet the URM label includes black, Latin@,1 Native American, Hawaiian and Pacific Islander students, and the majority designation is often both white and Asian students. The consequence of clustering leads to conclusions that are too generalized; for example, that black students will respond in a similar way to a treatment as do Latin@ students (Carpenter et al., 2006). Yet the different racial and ethnic groups that are included in the URM designation have very different cultures, histories, and exposure to college culture that could impact whether a particular classroom strategy is effective for them (Delpit, 2006). National trends in K–12 education, revealing different achievement patterns and trajectories for black and Latin@ students, also challenge the assumption that URM students are a homogeneous group (Reardon and Galindo, 2009).

To our knowledge, only two college-level curricular interventions in STEM, and none in biology, have subdivided the URM category into more fine-grained groups to explore the effectiveness of classroom interventions for these different student populations. In these studies, students of different racial/ethnic groups responded differently to the classroom interventions (Etkina et al., 1999; Beichner et al., 2007). This was demonstrated most dramatically by Beichner et al. (2007), in whose study white and black students were the only groups to benefit significantly from an active-learning intervention. These findings highlight the need for more studies to analyze college course performance by racial/ethnic groups. These smaller categories can still be problematic, as they still combine students with very different cultural backgrounds and experiences into broad categories such as white, Asian, Native American, and Latin@ students. This term is being increasingly used in the Latin@ community including many national organizations.

---

1Latin@ is a gender inclusive way of describing people of Latin American descent (Demby, 2013). The term is being increasingly used in the Latin@ community including many national organizations.
backgrounds that may be at odds with the middle-class cultural norms of universities (e.g., the emphasis on abstract over practical knowledge and independence over interdependence; Stephens et al., 2012; Wilson and Kittleson, 2013). The differences between first- and continuing-generation students have been shown to change how they respond to “best-practices” in teaching at the college level, sometimes to the extent that they respond oppositionally (Padgett et al., 2012). In biology, we are not aware of any studies that have explored the response of this population to an active-learning intervention, although there has been promising work with a psychology intervention (Harackiewicz et al., 2014).

In our study, we explored whether racial (black, white, Native American, Asian) and/or ethnic (Latin@) identity and first-generation versus continuing-generation status influenced a student’s response to the increased course structure. We hypothesized that different student groups would vary in the extent to which an active-learning intervention would influence their exam performance.

**How Do Active-Learning Interventions Change Course-Related Behaviors and Attitudes of Students?**

Understanding how interventions change course-related behaviors and attitudes is an important next step in education research, as these behaviors and attitudes mediate how the course structure influences performance (Singer et al., 2012). Some work has already described how active learning increases achievement at the college level, although this work is lacking in the STEM disciplines and usually only looks at the student body as a whole. Courses with more active learning are positively correlated with increased student self-reported motivation and self-efficacy (van Wyk, 2012) and a deeper approach to learning (Eley, 1992). Unfortunately, this work is only done in active-learning classrooms, and either there is no control group (cf. Keeler and Steinhorst, 1995; Cavanagh, 2011) or the study asks students to compare their experience with a different course with a different instructor and content in which they are currently enrolled (cf. Sharma et al., 2005). In our study, we examine how student attitudes and course-related behaviors change between a traditionally taught and an increased-structure course with the same content and instructor.

Reviewing the elements of successful classroom interventions suggests possible factors that could contribute to the increase in student achievement. For example, the increased course structure intervention involves the addition of three elements: graded preparatory assignments, extensive student in-class engagement, and graded review assignments (Table 1). Proponents of the increased course structure intervention have hypothesized that the additional practice led to the rise in student performance (Freeman et al., 2011). Yet providing opportunities for practice might not be enough. When and what students practice, as well as the context of and their perceptions of the practice may influence the impact of the extra practice on learning.

There are many possible factors that change with the implementation of increased course structure. We focus on three candidate factors, but it is important to recognize that these factors are not mutually exclusive or exhaustive.

**Factor 1. Time allocation:** Increasing course structure will encourage students to spend more time each week on the course, particularly on preparation. How students allocate their out-of-class study time can greatly influence their learning and course achievement. Many students adopt the strategy of massing their study time and cramming just before exams (Michaels and Miethe, 1989; McIntyre and Munson, 2008). Yet distributed practice is a more effective method for learning, particularly for long-term retention of knowledge (Dunlosky et al., 2013). The increased course structure helps students distribute their study time for the class by assigning daily or weekly preparatory and review assignments. These assignments 1) spread out the time students spend on the course throughout the quarter (distributed practice, rather than cramming just before exams) and 2) encourage students to engage with a topic before class (preparatory assignment) and then again in class (in-class activities) and again after class (review assignments). In addition, the preparatory assignments not only encourage students to read the book before class, but also have students answer questions related to the reading, which is a more effective method for learning new material than simply highlighting a text (Dunlosky et al., 2013). We believe that the outside assignments scaffold how students spend time on the course and are one of the primary factors by which increased course structure impacts student performance. However, this idea has never been explicitly tested. In this study, we asked students to report how much time they spent outside of class on the course weekly and what they spent that time doing. We predicted that students would spend more time each week on the course and would spend more time on the parts associated with course points. These results would imply an increase in distributed practice and demonstrate that the instructor can successfully guide what students spend time on outside of class.

**Factor 2. Classroom culture:** Increasing course structure will encourage students to perceive the class as a community. To learn, students must feel comfortable enough to be

<table>
<thead>
<tr>
<th>Table 1. The elements of a low-, moderate-, and high-structure course</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Graded preparatory assignments</strong> (example: reading quiz)</td>
</tr>
<tr>
<td>Low (traditional lecture) None or &lt;1 per week</td>
</tr>
<tr>
<td>Moderate Optional: 1 per week ≥1 per week</td>
</tr>
<tr>
<td>High</td>
</tr>
</tbody>
</table>

*aNeed either a preparatory or review assignment once per week, but not both.*
willing to take risks and engage in challenging thinking and problem solving (Ellis, 2004). High-stakes competitive classrooms dominated by a few student voices are not environments in which many students feel safe taking risks to learn (Johnson, 2007). The increased-structure format has students work in small groups, which may help students develop a more collaborative sense of the classroom. Collaborative learning in college has been shown to increase a sense of social support in the classroom as well as the sense that students like each other (Johnson et al., 1998). This more interdependent environment also decreases anxiety and leads to increased participation in class (Fassinger, 2000) and critical thinking (Tsui, 2002). Increased participation in in-class practice alone could lead to increased performance on exams. In addition, a more interdependent environment has been shown to be particularly important for the performance of first-generation students and Mexican students (Stephens et al., 2012; Savani et al., 2013). Finally, feeling like they are part of a community increases both performance and motivation, especially for historically underrepresented groups (Walton and Cohen, 2007; Walton et al., 2012). We predicted that students in an increased-structure course would change how they viewed the classroom, specifically, that they would feel an increased sense of community relative to students in low-structure courses.

Factor 3. Course value: Increasing course structure will increase the perceived value of the course to students. In the increased-structure course, students come to class having read the book, or at least worked through the preparatory assignment, and thus have begun the knowledge acquisition stage of learning. This shift of content acquisition from in class to before class opens up time in the classroom for the instructor to help students develop higher-order cognitive skills (Freeman et al., 2011), providing opportunities to encourage students to make connections between course content and real-world impacts and to work through challenging problems. These opportunities for practice and real-world connections are thought to be more engaging to students than traditional lecture (Handelsman et al., 2006). Thus, through increased engagement with the material (because of increased interest in it) student performance will increase (Carini et al., 2006). We predicted students in the increased-structure course would feel more engaged by the material and thus would value the course more.

We considered these three factors—time allocation, classroom culture, and course value—when surveying students about their perceptions and behaviors. We analyzed student survey responses in both the traditional and increased-structure course to identify patterns in responses that support the impact of these three factors on student performance.

In summary, we test the transferability of one active-learning intervention (increased course structure; Freeman et al., 2011) into a novel educational context. We expand upon the initial studies by 1) disaggregating student performance to test the hypothesis that student subpopulations respond differently to educational interventions and 2) using student self-reported data to identify possible factors (time allocation, classroom culture, course value) through which the intervention could be influencing student achievement.

METHODS AND RESULTS

The Course and the Students

The course, offered at a large research institution in the Southeast that qualifies as a more selective, full-time, 4-yr institution with a low transfer-in rate on the Carnegie scale, is a one-semester general introduction to biology serving a mixed-majors student population. The course is offered in both Fall and Spring semesters. Course topics include general introductions to the nature of science, cell biology, genetics, evolution and ecology, and animal physiology. The class met three times a week for 50 min each period. An optional laboratory course is associated with the lecture course, but lab grades are not linked to lecture grade. Although multiple instructors teach this course in a year, the data used in this study all come from six terms taught by the same instructor (K.A.H.). The instructor holds a PhD in pathology and laboratory medicine and had 6 yr of experience teaching this course before any of the terms used in this study.

The majority of students enrolled in the course were in their first year of college (69%), but the course is open to all students. The class size for each of the six terms of the study averaged 393 students. The most common majors in the course include biology, exercise and sports science, and psychology. The combined student demographics in this course during the years of this study were: 59% white, 13.9% black, 10.3% Latin@, 7.4% Asian, 1.1% Native American, and 8% of either undeclared race, mixed descent, or international origin. In addition, 66.3% of the students identified as female, 32.1% male, and 1.6% unspecified gender, and 24% of these students were first-generation college students.

The Intervention: Increasing Course Structure

Throughout our analyses, we compared the same course during three terms of low structure and three terms of moderate structure (Table 1). How these designations—low and moderate—were determined is explained later in the section Determining the Structure Level of the Intervention.

During the low-structure terms of this study (Spring 2009, Fall 2009, Spring 2010), the course was taught in a traditional lecture format in which students participated very little in class. In addition, only three homework assignments were completed outside the classroom to help students prepare for four high-stakes exams (three semester exams and one cumulative final).

In the reformed terms (Fall 2010, Spring 2011, Fall 2011), a moderate-structure format was used with both in-class and out of class activities added. The elements added—guided-reading questions, preparatory homework, and in-class activities—are detailed below, and Table 2 gives some specific examples across one topic.

Guided-Reading Questions. Twice a week, students were given ungraded, instructor-designed guided-reading questions to complete while reading their textbook before class. These questions helped to teach active reading (beyond highlighting) and to reinforce practice study skills, such as drawing, using the content in each chapter (Table 2; Supplemental Material, section 1). While these were not graded, the expectation set by the instructor was that the daily activities built from this content and referred to them, without covering them in the same format. Keys were not posted.
Table 2. Sample question types associated with the three assignment types added during the moderate-structure terms

<table>
<thead>
<tr>
<th>Preclass (ungraded)</th>
<th>Preclass (graded)</th>
<th>In-class (extra credit)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example guided-reading questions</td>
<td>Example preparatory homework question</td>
<td>Example in-class questions</td>
</tr>
<tr>
<td>1. Examine Figure 8.14, why are the chromosomes colored red and blue in this figure? What does red or blue represent? 2. Describe in words and draw how independent orientation of homologues at metaphase I produces variation.</td>
<td>Independent orientation of chromosomes at metaphase I results in an increase in the number of: a) Sex chromosomes b) Homologous chromosomes c) Points of crossing over d) Possible combinations of characteristics e) Gametes</td>
<td>Students were shown an illustration of a diploid cell in metaphase I with the genotype AaBbDd. For all questions, students were told to “ignore crossing over.” 1. For this cell, what is n = ? 2. How many unique gametes can form? That is, how many unique combinations of chromosomes can form? 3. How many different ways in total can we draw metaphase I for this cell? 4. How many different combinations of chromosomes can you make in one of your gametes?</td>
</tr>
</tbody>
</table>

Preparatory Homework. Students were required to complete online graded homework associated with assigned readings before coming to class (Mastering Biology for Pearson’s *Campbell Biology: Concepts and Connections*). The instructor used settings for the program to coach the students and help them assess their own knowledge before class. Students were given multiple opportunities to answer each question (between two and six attempts, depending on question structure) and were allowed to access hints and immediate correct/incorrect answer feedback. The questions were typically at the knowledge and comprehension levels in Bloom’s taxonomy (Table 2).

In-Class Activities. As course content previously covered by lecture was moved into the guided-reading questions and preparatory homework, on average 34.5% of each class session was now devoted to activities that reinforced major concepts, study skills, and higher-order thinking skills. Students often worked in informal groups, answering questions similar to exam questions by using classroom-response software (www.polleverywhere.com) on their laptops and cell phones. Thirty-six percent of these questions required a student to apply higher-order cognitive skills such as application of concepts to novel scenarios or analysis (see Supplemental Material, section 2, for methods). Although responses to in-class questions were not graded, students received 1–2 percentage points of extra credit on each of four exams if they participated in a defined number of in-class questions. The remaining 65.5% of class time involved the instructor setting up the activities, delivering content, and course logistics. These percentages are based on the observation of videos from four randomly chosen class session videos. The course was videotaped routinely, so the instructor did not know in advance which class sessions would be scored.

Determining the Structure Level of the Intervention

Using the data from two articles by Freeman and colleagues (Freeman et al., 2007, 2011) and consulting with Scott Freeman (personal communication) and the Biology Education Research Group at the University of Washington, we identified the critical elements of low, moderate, and high structure (Table 1). Based on these elements, our intervention was a “moderate” structure course: we had weekly graded preparatory homework, students were talking on average 35% of class time, and there were no graded review assignments.

Study 1: Does the Increased Course Structure Intervention Transfer to a Novel Environment?

Total Exam Points by Course Structure. Our measure of achievement was total exam points. We chose this measure over final grade, because the six terms of this course differed in the total points coming from homework (3 vs. 10%) and the opportunity for bonus points could inflate the final grade in the reformed class. Instead, we compared the total exam points earned out of the possible exam points. As total exam points varied across the six terms by 5 points (145–150), all terms were scaled to be out of 145 points in the final data set. Determining the Structure Level of the Intervention

Using the data from two articles by Freeman and colleagues (Freeman et al., 2007, 2011) and consulting with Scott Freeman (personal communication) and the Biology Education Research Group at the University of Washington, we identified the critical elements of low, moderate, and high structure (Table 1). Based on these elements, our intervention was a “moderate” structure course: we had weekly graded preparatory homework, students were talking on average 35% of class time, and there were no graded review assignments.

Study 1: Does the Increased Course Structure Intervention Transfer to a Novel Environment?

Total Exam Points by Course Structure. Our measure of achievement was total exam points. We chose this measure over final grade, because the six terms of this course differed in the total points coming from homework (3 vs. 10%) and the opportunity for bonus points could inflate the final grade in the reformed class. Instead, we compared the total exam points earned out of the possible exam points. As total exam points varied across the six terms by 5 points (145–150), all terms were scaled to be out of 145 points in the final data set. As this study took place over 4 years, we were concerned that term-to-term variation in student academic ability and exam difficulty could confound our survey and achievement results. To be confident that any gains we observed were due to the intervention and not these other sources of variation, we controlled for both exam cognitive level (cf. Crowe et al., 2008) and student prior academic achievement (for more details see Supplemental Material, section 2). We found that exams were similar across all six terms and that the best control for prior academic achievement was a student’s combined SAT math and SAT verbal score (Table 3; Supplemental Material, section 2). We therefore used SAT scores as a control for student-level variation in our analyses and did not further control for exams.

Course and Exam Failure Rates by Course Structure. To become a biology major, students must earn a minimum of a “C−” in this course. Thus, for the purpose of this study, we considered a grade below 72.9% to be failing, because the
Model 3: Impact of course structure on student performance

	Student performance influenced by course structure

	Model 2: Impact of course structure on student performance

	Model 3: Impact of course structure on student performance


\[
\text{Outcome} \sim \text{Term} + \text{Combined SAT scores} + \text{Gender} + \text{Course Structure}
\]

\[
\text{Outcome} \sim \text{Term} + \text{SAT scores} + \text{Gender} + \text{Course Structure} + \text{Race} + \text{Term} \times \text{Course Structure}
\]

\[
\text{Outcome} \sim \text{Term} + \text{SAT scores} + \text{Gender} + \text{Course Structure} + \text{First-generation} + \text{First-generation} \times \text{Course Structure}
\]

\(^{a}\)Bolded terms in models 2 and 3 are the new additions that test the specific hypotheses that the impact of course structure will vary by student populations. The outcome variable is either student achievement on exams or student failure rates.

Student earning this would not be able to move on to the next biology course. We measured failure rates in two ways: 1) final grade and 2) total exam points. Although the components contributing to final course grade changed across the study, this “C−” cutoff for entering the biology major remained consistent. This measure may be more pertinent to students than overall exam performance, because it determines whether or not they can continue in the major.

To look more closely at whether increased student learning was occurring due to the intervention, we looked at failure rates on the exams themselves. This measure avoids the correlation of any boost in performance due to extra credit or homework points or deviations from a traditional grading scale but is not as pertinent to retention in the major as course grade.

The statistical analysis for this study is paired with that of study 2 and is described later.

**Study 2. The Effectiveness of Increased Course Structure Across Different Student Populations?**

In addition to identifying whether an overall increase in achievement occurred during the moderate-structure terms, we included categorical variables in our analyses to determine whether student subpopulations respond differently to the treatment. We focused on two designations: 1) student ethnic, racial, or national origin, which included the designations of Asian American, black, Latin@, mixed race/ethnicity, Native American, white, and international students; and 2) student generational status (first-generation vs. continuing-generation college student). Both of these factors were determined from student self-reported data from an in-class survey collected at the end of the term.

**Statistical Analyses: Studies 1 and 2**

**Total Exam Points Earned by Course Structure and Student Populations.** We modeled total exam points as continuous response and used a linear regression model to determine whether moderate course structure was correlated with increased exam performance (Table 3). In our baseline model, we included student combined SAT scores, gender identity (in this case, a binary factor: 0 = male, 1 = female), and the term a student was in the course (Fall vs. Spring) as control variables. Term was included, because the instructor has historically observed that students in the Spring term perform better than students in the Fall term.

To test our first hypothesis, that increasing the course structure would increase performance (study 1), we included treatment (0 = low structure, 1 = moderate structure) as a binary explanatory variable. To test our second hypothesis, that students from different populations may differ in their response to the classroom intervention, we ran two models (Table 3) that included the four variables described above and either 1) student racial and ethnic group (a seven-level factor) or 2) student first-generation status (a binary factor: 1 = first generation, 0 = continuing generation). If any of these demographic descriptors were not available for a student, that student was not included in the study.

We ran separate regression models for race/ethnicity and generation status, because we found these terms were correlated in an initial test of correlations between our possible explanatory variables (Kruskal-Wallis \( \chi^2 = 68.1, df = 5, p < 0.0001 \)). Thus, to avoid any confounds due to multicollinearity (correlation between two explanatory variables), we ran each term in a separate model.

**Course and Exam Failure Rates for Student Populations.** We also explored whether the failure rate in the course decreased with the implementation of moderate course structure and whether different populations of students responded to the treatment differently. Our response variable was either 1) passing or failing the class or 2) passing or failing the exams (with \( p < 0.0001 \)). Thus, to avoid any confounds due to multicollinearity (correlation between two explanatory variables), we ran each term in a separate model.

**Results: Studies 1 and 2**

**Total Exam Points Earned by Student Populations: Performance Increased for All Students but Increased Disproportionately for Black and First-Generation Students**

Exam performance by course structure and student race/ethnicity/nationality. In the low-structure terms—after we accounted for differences in SAT math and reading scores, gender identity, and differences between term—Asian, Native American, and white students had the highest achievement. Black (\( \beta = -8.1 \pm 1.6 SE, p < 0.0001 \)) and Latin@ (\( \beta = -3.4 \pm 1.7 SE, p = 0.044 \)) students scored significantly fewer exam points (6 and 2% fewer points, respectively; Table 4). It is important to note that the Native American category in our analysis contains very few students, and these results may be due to a lack of statistical power, rather than there being no real difference between white students and Native American student performance.

In the moderate-structure term, after we controlled for SAT scores, gender identity, and term, the classroom intervention
Table 4. Students of all racial/ethnic/national populations benefit from increased course structure and black students benefit disproportionatelya

<table>
<thead>
<tr>
<th>Regression coefficients</th>
<th>Estimate ± SE</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model intercept</td>
<td>4.6 ± 4.52</td>
<td>0.310</td>
</tr>
</tbody>
</table>

Exam performance patterns under low structure

Race/Ethnicity/Nationality:

- Native American: $-2.4 ± 4.38$, p = 0.569
- Asian: $0.1 ± 2.19$, p = 0.951
- Black: $-8.10 ± 1.56$, p = 0.0001
- Latin@: $-3.4 ± 1.67$, p = 0.044
- Mixed Race: $0.8 ± 3.85$, p = 0.826
- International: $-7.0 ± 4.91$, p = 0.157

Exam performance patterns under moderate structure

- Moderate Structure: $4.6 ± 1.01$, p = 0.0001

Class Structure*Race/Ethn./Nat.:

- Moderate*Native American: $-2.3 ± 6.56$, p = 0.726
- Moderate*Asian: $0.2 ± 2.74$, p = 0.948
- Moderate*Black: $4.5 ± 2.08$, p = 0.031
- Moderate*Latin@: $2.4 ± 2.40$, p = 0.317
- Moderate*Mix Race: $-2.1 ± 4.66$, p = 0.657
- Moderate*International: $5.4 ± 6.96$, p = 0.440

Controls for student characteristics and term

- Term: (reference level: Fall)
  - Spring: $4.0 ± 0.77$, p = 0.0001
- SAT.Combined: $0.08 ± 0.003$, p = 0.0001
- Gender: (reference level: Male)
  - Female: $1.8 ± 1.13$, p = 0.022

Table 5. All students benefit from increased course structure and first-generation students benefit disproportionatelya

<table>
<thead>
<tr>
<th>Regression coefficients</th>
<th>Estimate ± SE</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model intercept</td>
<td>$-6.1 ± 4.00$</td>
<td>0.128</td>
</tr>
</tbody>
</table>

Exam performance patterns under low structure

- First-generation: $-3.9 ± 1.19$, p = 0.0012

Exam performance patterns under moderate structure

- Moderate: $5.4 ± 0.87$, p = 0.003
- Moderate*First Gen: $3.5 ± 1.64$, p = 0.032

Controls for student characteristics and term

- SAT.Combined: $0.08 ± 0.003$, p = 0.0001
- Gender: (reference level: Male)
  - Female: $1.6 ± 0.78$, p = 0.680
- Term: (reference level: Fall)
  - Spring: $4.1 ± 0.76$, p = 0.0001

aCoefficients from regression model of impact of race and course structure on exam points earned (out of 145). Coefficient estimates are in terms of raw exam points. The categorical variable Race/Ethnicity/Nationality represents achievement by racial group under low structure relative to the achievement of white students under low structure. The Class Structure term represents the gain all students see in the moderate-structure course. The interaction term between class structure and generation status (Class Struct*Gen. Status) represents the gains first-generation students experience relative to the gains of continuing-generation students under moderate structure. Significant terms for Course Structure*Gen. Status indicate a disproportionate impact of moderate structure on first-generation students.

We controlled for term, SAT math and reading scores, and first-generation status ($\beta = 1.6 ± 0.78$ SE, p = 0.068; Table 5). Second, increasing the course structure provided all students, regardless of first-generation status, a slightly larger boost (3.7%, $\beta = 5.4 ± 0.87$ SE, p = 0.003; Table 5).

Our main focus, however, is whether first-generation students responded differently to increased course structure relative to continuing-generation students. Under low structure, there was a 2.5% difference in exam points earned between first-generation students and continuing-generation students ($\beta = -3.9 ± 1.2$ SE, p = 0.001; Table 5). With increased course structure, the performance of all students increased by 3.7%, and first-generation students experienced an additional 2.4% increase in exam performance for a total 6.1% increase ($\beta = 3.5 ± 1.6$ SE, p = 0.032; Table 5). This disproportionate increase in first-generation student performance closes the achievement gap between first- and continuing-generation students.

Overall, the major pattern associated with student achievement was that exam points earned increased under moderate structure relative to low structure and increased disproportionately for black and first-generation students (Figure 1).
Course and Exam Failure Rates: Failure Rates Decreased for All Students

Raw failure rates in the course also dropped when comparing low with moderate structure. Overall, without controlling for any aspects of the course or student ability, 26.6% of the class earned a course grade of a “C−” or lower during the low-structure terms. With moderate structure, that failure rate dropped by 41.3% to 15.6%. Exam failure (earning <72.9% of the possible exam points) also decreased by 17.3% under moderate structure (from 52.6 to 43.5%).

After we controlled for student combined SAT score and the term, students experiencing moderate structure were 2.3 times more likely to earn above a “C−” in the course ($\beta = 0.828 \pm 0.19 SE, p < 0.0001$) and were 1.6 times more likely to earn more than 72.9% of the possible exam points ($\beta = 0.486 \pm 0.130 SE, p = 0.0001$). There was not a significant interaction between student race, gender, or first-generation status and course structure on this coarse scale.

Study 3: What Factors Might Influence Student Achievement in the Course with Increased Structure?

Methods. A 30-question survey was given to students immediately after the final exam in one term of the low-structure course and in three terms of the moderate-structure course. The survey questions focused on student course-related behaviors (Likert scale), student perception of the course, (Likert scale), and self-reported demographic variables (Supplemental Material, section 4). Specifically, we were interested examining three aspects of increased structure that might influence student learning. Survey questions had four or five response options. The questions associated with each factor are detailed in the following sections.

Factor 1. Time Allocation. We predicted that increased course structure would not only increase the time students spent on class material outside of class each week (thus distributing their practice throughout the term) but also the time they spent on behaviors associated with graded assignments. To test this prediction, we had students in both the low- and moderate-structure courses report how many hours they spent studying a week and the frequency of behaviors related to preparing for and reviewing after class (Table 6, Supplemental Material, section 4). With our survey questions, it is not possible to parse out whether the increased hours of practice each week also led to an overall increase in the amount of time students spent practicing.

Factor 2. Classroom Culture. We predicted that students would feel a greater sense of interdependence in the moderate-structure course. To test this, we asked students to report how frequently they participated in class and how frequently they studied in groups outside of class. We also asked them three questions related to how interdependent they perceived the students in the class to be: how well they thought students in the class knew each other, if they believed students in the class tried to help one another, and whether they felt the class was a community (Table 6; Supplemental Material, section 4).

Factor 3. Course Value. We predicted that students would value the class and skills they acquired through the class more under moderate structure, because more higher-order skills were incorporated into the class. An assumption of this prediction is that students recognized that this class required higher-order thinking. To test this assumption, we asked students in both the low- and moderate-structure courses to identify how much of the course involved memorization. To assess the value students place in the course, we asked them to report lecture attendance, the importance of lecture for their learning, and the usefulness of the skills learned in the course for their future classes (Table 6; Supplemental Material, section 4).

In addition to looking at general impacts of active learning, we explored whether there were differences between student populations in their survey responses. These differences could help us understand why some student populations benefit more than others from the increased course structure intervention. The populations we focused on were identified through study 1: black and first-generation students performed disproportionately better in the increased-structure course relative to other student populations.
Table 6. Changes in student behaviors and perceptions with increased course structure

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Low-structure term (raw median response)</th>
<th>Moderate-structure terms (raw median response)</th>
<th>Odds ratio: likelihood to increase with moderate structure (95% CI)</th>
<th>Odds ratio: likelihood to increase in SAT scores (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Factor 1. Time allocation:</strong> Increasing course structure will encourage students to spend more time each week on the course, particularly on preparation.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hours spent studying/week&lt;br&gt;(0, 1–3, 4–7, 7–10, &gt;10 h)</td>
<td>1–3 h</td>
<td>4–7 h</td>
<td>2.60 (2.02–3.35)</td>
<td>0.982 (0.974–0.990)</td>
</tr>
<tr>
<td>Complete readings before class&lt;br&gt;(Never, Rarely, Sometimes, Often)</td>
<td>Rarely</td>
<td>Sometimes</td>
<td>1.97 (1.54–2.52)</td>
<td>0.994 (0.985–1.00)</td>
</tr>
<tr>
<td>Preparatory homework importance&lt;br&gt;(Not at all, Somewhat, Important, Very)</td>
<td>Somewhat</td>
<td>Important</td>
<td>4.6 (3.56–5.85)</td>
<td>0.98 (0.97–0.98)</td>
</tr>
<tr>
<td>Review notes after class&lt;br&gt;(Never, Rarely, Sometimes, Often)</td>
<td>Sometimes</td>
<td>Sometimes</td>
<td>0.738 (0.583–0.933)</td>
<td>0.972 (0.965–0.980)</td>
</tr>
<tr>
<td>Complete textbook review questions&lt;br&gt;(Never, Rarely, Sometimes, Often)</td>
<td>Rarely</td>
<td>Rarely</td>
<td>0.50 (0.400–0.645)</td>
<td>0.98 (0.972–0.99)</td>
</tr>
<tr>
<td><strong>Factor 2. Classroom culture:</strong> Increasing course structure will encourage students to perceive the class as more of a community.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contribute to classroom discussions&lt;br&gt;(Never, Rarely, Sometimes, Often)</td>
<td>Never</td>
<td>Rarely</td>
<td>1.13 (0.890–1.44)</td>
<td>0.99 (0.988–1.00)</td>
</tr>
<tr>
<td>Work with a classmate outside of class&lt;br&gt;(Never, Rarely, Sometimes, Often)</td>
<td>Sometimes</td>
<td>Sometimes</td>
<td>0.83 (0.664–1.06)</td>
<td>0.984 (0.977–0.991)</td>
</tr>
<tr>
<td>Believe students in class know each other&lt;br&gt;(Strongly disagree, Disagree, Neutral, Agree, Strongly agree)</td>
<td>Neutral</td>
<td>Neutral</td>
<td>2.4 (1.92–3.09)</td>
<td>0.996 (0.989–1.00)</td>
</tr>
<tr>
<td>Believe students in class help each other&lt;br&gt;(Strongly disagree, Disagree, Neutral, Agree, Strongly agree)</td>
<td>Agree</td>
<td>Agree</td>
<td>1.22 (0.948–1.57)</td>
<td>1.01 (0.999–1.02)</td>
</tr>
<tr>
<td>Perceive class as a community&lt;br&gt;(Strongly disagree, Disagree, Neutral, Agree, Strongly agree)</td>
<td>Neutral</td>
<td>Neutral</td>
<td>1.99 (1.57–2.52)</td>
<td>0.986 (0.979–0.993)</td>
</tr>
<tr>
<td><strong>Factor 3. Course value:</strong> Increasing course structure will increase the value of the course to students.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amount of memorization&lt;br&gt;(Most, Quite a bit, Some, Very Little, None)</td>
<td>Some</td>
<td>Some</td>
<td>1.07 (0.84–1.35)</td>
<td>0.98 (0.982–0.997)</td>
</tr>
<tr>
<td>Attend lecture&lt;br&gt;(Never, Rarely, Sometimes, Often)</td>
<td>Often</td>
<td>Often</td>
<td>0.72 (0.471–1.09)</td>
<td>0.984 (0.971–0.997)</td>
</tr>
<tr>
<td>Use of skills learned&lt;br&gt;(Strongly disagree, Disagree, Neutral, Agree, Strongly agree)</td>
<td>Agree</td>
<td>Agree</td>
<td>0.909 (0.720–1.15)</td>
<td>0.991 (0.983–0.998)</td>
</tr>
<tr>
<td>Lecture importance&lt;br&gt;(Not at all, Somewhat, Important, Very)</td>
<td>Very Important</td>
<td>Important</td>
<td>0.57 (0.448–0.730)</td>
<td>0.998 (0.991–1.01)</td>
</tr>
</tbody>
</table>

*aThe second and third columns are the raw median responses under each structure. The fourth and fifth columns are the odds ratios from the log-odds regression including course structure and SAT scores as explanatory variables (>1 = students more likely to report a higher value; <1 = students more likely to report a lower value). SAT scores were divided by 10 to increase size of regression coefficients, so confidence intervals would be interpretable. Bolded odds ratios are significant.

Statistical Analysis: Study 3

General Patterns. We compared student responses to 14 survey questions concerning student course-related behaviors and perceptions of the classroom environment between one term of low course structure and three terms of moderate structure. Survey responses were ordered categorical responses (with four to five levels per question), so we used proportional log-odds regression models (implemented with the MASS package in R; Venables and Ripley, 2002). The proportional log-odds model works well for tightly bounded or ordinal data. The model compares the levels in the response variable by running a series of dichotomous comparisons (“never” vs. “rarely,” “rarely” vs. “sometimes,” etc.). Thus, the output of the log-odds regression is the effect that a change in the explanatory variable (e.g., presence or absence of the classroom intervention) has on the odds that a student will report a higher rather than a lower response (i.e., “strongly agree” vs. “agree”) averaged across all possible levels of response (Antoine and Harrell, 2000).

Predictor variables for each model included: SAT combined score and course structure. We included SAT scores as a predictor, because we believed that students with higher levels of academic preparedness might differ from those with lower preparedness in terms of their course-related behaviors and attitudes. Thus, the model used for each question was: Survey response = Intercept + β*SATI.Comb + β*Treat. p Values were adjusted to account for false discovery rates due to multiple comparisons (Pike, 2011).
Population-Specific Patterns. On the basis of study 1, we identified two target populations who responded most strongly to the increased course structure: black and first-generation students. To test whether these two groups responded differently to any of the survey questions, we used forward stepwise model selection in the sequence that follows. We started with the base model, SAT math and verbal scores and treatment (Survey response = Intercept + β*SATI.Comb + β*Treat), and then added a main effect of first-generation status or black racial identity (i.e., Survey response = Intercept + β*SATI.Comb + β*Treat + β*First.Gen). We compared these two models to determine whether adding the new variable significantly increased the fit of the model to the data using standard model selection techniques. If the model with the population variable was significant, we then added an interaction between treatment and that variable (Survey response = Intercept + β*SATI.Comb + β*Treat + β*First.Gen*Treat) and tested the fit of this new model to the data. The significant difference in model fit (if present) was calculated using a type II analysis of variance (implemented with the car package in R; Fox and Weisberg, 2010). Again, p values were adjusted to account for false discovery rates due to multiple comparisons (Pike, 2011).

Correlation of Student Study Strategies/Perceptions with Total Exam Points. We also used survey responses to determine how student behaviors and attitudes correlated with their exam scores and whether this relationship with exam points was mediated by course structure.

Initially, we used the gamma rank correlation to explore whether responses on any of the 14 survey questions were correlated with responses on the other questions. This analysis was implemented in R using the rococo package (Bodenhofer and Klavon, 2008). Even after correcting for the false discovery rate due to multiple comparisons, there were many moderate correlations between response in one survey question and response in another possibly leading to multicollinearity (for specific results, see Supplemental Material, section 3).

To be conservative, we ran 14 regression models, each linking a single survey question, the control for prior student academic achievement, and course structure to exam performance. This allowed us to determine whether the behavior or attitude in question was correlated with exam performance and whether this relationship was mediated by classroom structure. It did not allow us to identify which behaviors and attitudes relative to each other were most important.

Results: Study 3

General Patterns: Student Behaviors and Perceptions Changed with Increased Course Structure. Students reported employing different study strategies and perceiving the components of the course differently in low- and moderate-structure courses (Table 6).

Factor 1. Time Allocation. All five of the questions related to time allocation varied significantly with course structure (Table 6). As predicted, students spent more time each week preparing for class in the moderate-structure course. Specifically, students were 2.6 times more likely to report spending more hours a week studying for biology ($\beta = 0.95 \pm 0.13 \text{ SE}, p < 0.0001$). Students focused more on preparing for class in the moderate-structure course versus the low-structure course: after we controlled for SAT math and reading scores, students were 2.0 times as likely to complete reading assignments before class ($\beta = 0.68 \pm 0.12 \text{ SE}, p = 0.0001$) and were 4.7 times more likely to report that the homework assignments were important for their understanding of course material in the moderate-structure term ($\beta = 1.5 \pm 0.12 \text{ SE}, p > 0.0001$). Interestingly, even with the additional investment of hours each week, a focus on preparation seemed to represent a trade-off with time spent reviewing: after we controlled for SAT math and reading scores, students were 1.4 times less likely to complete their notes after class as frequently ($\beta = -0.30 \pm 0.12 \text{ SE}, p = 0.011$) and 1.9 times less likely to complete the practice questions at the end of each book chapter ($\beta = -0.68 \pm 0.12 \text{ SE}, p < 0.0001$).

Factor 2. Classroom Culture. Of the five questions focused on the class climate, only two changed significantly with course structure (Table 6). As predicted, student in the moderate-structure terms were 2.0 times more likely to report a stronger sense of classroom community ($\beta = 0.69 \pm 0.12 \text{ SE}, p < 0.0001$) and 2.4 times more likely to agree with the statement that “students in this class know each other” ($\beta = 0.89 \pm 0.12 \text{ SE}, p < 0.0001$). The other three outcomes, which we expected would increase but did not, were: how strongly students believed that students in the class helped one another ($\beta = 0.20 \pm 0.13 \text{ SE}, p = 0.12$), the frequency at which students worked with a partner outside of class ($\beta = 0.18 \pm 0.12 \text{ SE}, p = 0.14$), and the frequency at which students participated in class ($\beta = 0.12 \pm 0.12 \text{ SE}, p = 0.32$).

Factor 3. Course Value. Although we predicted outcomes that would suggest students valued the course more, we actually saw a decline (Table 6): students in the moderate-structure terms were 1.7 times less likely to rate the lecture component as important as students in the low-structure term ($\beta = -0.56 \pm 0.12 \text{ SE}, p > 0.0001$). After we controlled for SAT math and reading scores, students also did not vary in their frequency of lecture attendance (although this could be because it was high to begin with; $\beta = -0.32 \pm 0.21 \text{ SE}, p = 0.13$). Student perception of the importance of the skills they learned in the class did not vary between course structures ($\beta = 0.09 \pm 0.12 \text{ SE}, p = 0.42$) nor did they perceive that the moderate-structure course involved more cognitive skills other than memorization ($\beta = 0.07 \pm 0.12 \text{ SE}, p = 0.58$).

Population-Specific Patterns
Black Students Demonstrate Differences in Behaviors and Perceptions among Student Populations. On the basis of the results in study 1, which demonstrated that increased course structure was most effective for black and first-generation students, we explored student survey responses to determine whether we could document what was different for these populations of students.

We identified one behavior and three perception questions for which adding a binomial variable identifying whether a student was part of the black population or not increased the fit of the log-odds regression to the data. These differential responses may help us elucidate why this population responded so strongly to the increased-structure treatment.

The one behavior that changed disproportionately for black students relative to other students in the class was speaking in class. Under low structure, black students were 2.3 times more
The Increased-Structure Intervention Can Transfer across Different Instructors, Different Student Bodies, and Different Courses (Majors vs. Nonmajors)

One of the concerns of any classroom intervention is that the results depend on the instructor teaching the course (i.e., the intervention will work for only one person) and the students in it. We can test the independence of the intervention by replicating it with a different instructor and student body and measuring whether similar impacts on student achievement occur. The university at which this study took place is quite different from the university where the increased course structure intervention was developed (Freeman et al., 2011). Both universities are R1 institutions, but one is in the Southeast (and has a large black and Latin@ population), whereas the original university was in the Pacific Northwest (and has a high Asian population). Yet we find very similar results: in the original implementation of moderate structure in the Pacific Northwest course, the failure rate (defined as a course grade that would not allow a student to continue into the next course in the biology series) dropped from 18.2% to an average of 12.8% (a 29.7% reduction; Freeman et al., 2011). In our implementation of moderate structure, the failure rate dropped by a similar magnitude: from 26.6% to 15.6% (a 41.3% reduction). This result indicates that the impact of this increased-structure intervention may be independent of instructor and that the intervention could work with many different types of students.

Some Students Benefit More Than Others from Increased Course Structure

We found that transforming a classroom from low to moderate structure increased the exam performance of all students by 3.2%, and black students experienced an additional 3.1% increase (Figure 1A), and first-generation students experienced an additional 2.5% increase relative to continuing-generation students (Figure 1B). These results align with the small body of literature at the college level that indicates classroom interventions differ in the impact they have on student subpopulations (Kim, 2002; Preszler, 2009; Haak et al., 2011). Our study is novel in that we control for both student past academic achievement and disaggregate student racial/ethnic groups beyond the URM/non-URM binary. Our approach provides a more nuanced picture of how course structure impacts students of diverse demographic characteristics (independent of academic ability).

One of the most exciting aspects of our results is that we confirm that active-learning interventions influence the achievement of student subpopulations differentially. This finding is supported by both work in physics (Beichner et al., 2007), which found an intervention only worked for black and white students, and work in psychology, which revealed Asian-American students do not learn as well when they are told to talk through problems out loud (Kim, 2002). These studies highlight how important it is for us to disaggregate our results by student characteristics whenever possible, as overall positive results can mask actual differential outcomes present in the science classroom. Students come from a range of educational, cultural, and historical backgrounds and face different challenges in the classroom. It is not surprising that in the face of this diversity one intervention type does not fit all students equally.

DISCUSSION

With large foundation grants working toward improving STEM education, there has been a push for determining the transferability of specific educational innovations to “increase substantially the scale of these improvements within and across the higher education sector” (NSF, 2013). In this study, we provide evidence that one course intervention, increased course structure (Freeman et al., 2011), can be transferred from one university context to another. In addition to replicating the increase in student achievement across all students, we were able to elaborate on the results of prior research on increased course structure by 1) identifying which student populations benefited the most from the increased course structure and 2) beginning to tease out the factors that may lead to these increases.
Comparing our results with published studies in STEM focused on historically underrepresented groups, we see that our achievement results are of a similar magnitude to other interventions. Unlike our intervention, previous interventions generally are not implemented within an existing course but are either run as separate initiatives or separate courses or are associated with a series of courses (i.e., involved supplemental instruction [SI]; cf. Maton et al., 2000; Matsui et al., 2003). These SI programs are effective, but can be costly (Barlow and Villarejo, 2004), and because of the cost, they are often not sustainable. Of seven SI programs that report data on achievement and retention in the first term or first two terms of the program, and thus are directly comparable to our study results, failure rate reductions ranged from 36.3 to 77%, and achievement increased by 2.4–5.3% (Table 7). In our study, the failure rate reduction was 41.3%, and overall exam performance increased by 3.2% (6.2% for black students and 6.1% for first-generation students), which is within the range of variation for the short-term results of the SI studies. These short-term results may be an underestimate of the effectiveness of the SI programs, as some studies have shown that their effectiveness increases with time (Born et al., 2002). Yet the comparison still reveals promising results: one instructor in one course, without a large influx of money, can make a difference for students as large in magnitude as some supplemental instruction programs.

Exploring How Increased Course Structure Increases Student Performance

Survey data allowed us to explore how student course-related behaviors and attitudes changed with increased course structure. We focused on three specific factors and found evidence that changes in time allocation contributed to increased performance and some support for changes in classroom culture also impacting learning. We did not find evidence to support the idea that the value students found in the course influenced their performance.

Factor 1. Time Allocation. Under low structure, students on average spent only 1–3 h on the course outside of class, rarely came to class having read the assigned readings, and were highly dependent on the lecture for their learning. Students also placed little value on the occasional preparatory homework assignments. With the implementation of moderate structure, students increased the amount of time they spent on the course each week to 4–7 h, were twice as likely to come to class having read the assigned readings, and saw the preparatory assignments as being equally as important for their learning as the actual lecture component. These shifts in behaviors and perceptions support our hypothesis that increased course structure encourages students both to distribute their studying throughout the term and to spend more time on behaviors related to graded assignments.

We believe that these changes in student behaviors and perceptions occurred because of the structure of accountability built into the moderate-structure course. Students reading before class is an outcome almost all instructors desire (based on the ubiquitous syllabus reading lists), but it is evident from our study and others that, under low structure, students were on average “rarely” meeting this expectation (see also Burchfield and Sappington, 2000). We found the dual method of assigning preparatory homework and making the reading more approachable with ungraded guided-reading questions increased the frequency of students reading before class. It seemed that course points (accountability) were necessary to invoke this change in student behavior, because we did not see a similar increase in the frequency with which students reviewed notes after class. It is possible that moving to high structure (Freeman et al., 2011), with its weekly graded review assignments, could increase the achievement of our students even more, because they would be held accountable for reviewing their notes more frequently.

Factor 2. Classroom Culture. We found some evidence to support the hypothesis that increased course structure creates a community environment rather than a competitive environment. Under low structure, students did not seem to get to know the other students in the class and did not positively view the class as a community (although they did believe that students in the class tried to help one another). With increased structure, students were two times more likely to view the class as a community and 2.4 times more likely to say students in the class knew each other.

This result is a critical outcome of our study, arguably as important as increased performance, because a sense of being part of a community (belonging) is crucial for retention (Hurtado and Carter, 1997; Hoffman et al., 2002) and has been
correlated with increased performance for first-generation students (Stephens et al., 2012). When discussing reasons for leaving STEM, many students, particularly students of color and women, describe feelings of isolation and lack of belonging (Hewlett et al., 2008; Cheryan et al., 2009; Strayhorn, 2011). Because introductory courses are some of the first experiences students have in their major, these could potentially play a role in increasing retention simply by facilitating connections between students through small-group work in class.

**Factor 3. Course Value.** We did not find support for the hypothesis that students in the moderate-structure class found the course to be more valuable than students in the low-structure course. First, there was no difference in how much students valued the skills they learned in the course, but this could be because they did not recognize that the low- and moderate-structure terms were asking them to do different things. Across both terms, students on average believed that they were doing the same amount of memorizing versus higher-order skills such as application and analysis, even though the instructor emphasized higher-order skills more in the moderate-structure terms. In addition, behaviorally, we did not see any evidence of a higher value associated with the course in terms of increased attendance. In fact there was no difference in attendance across treatments. The attendance result was surprising to us, because increased attendance has been shown to be a common result of making a classroom more active (Caldwell, 2007; Freeman et al., 2007); however, these previous interventions all assigned course points to in-class participation, whereas our interventions only gave students bonus points for participation. In a comparison of in-class attendance with and without points assigned to class participation, Freeman et al. (2007) found that attendance dropped in the class in which no points were assigned. Thus, it is possible that attendance in these classes could be increased in the future if points rather than extra credit were assigned for participation. This idea is supported by our data that it is actually the students with the highest predicted achievement (i.e., highest SAT scores) who are more likely to miss lecture. Because these students already were doing well in the course, it may be that the motivation of receiving a few bonus points for attending class was not enough encouragement.

Additional evidence that changes in time allocation and classroom culture contribute to achievement comes from the correlation between survey responses and exam performance. Under moderate structure, the number of hours a student spent studying per week and a higher sense of community were both positively correlated with exam performance.

The support for these two factors, time allocation and classroom culture, helps us identify potential critical elements for the implementation of the increased-structure intervention. First, students need to be made accountable for preparing before attending class. This can take multiple forms, including guided-reading questions, homework, and/or reading quizzes before class or at the start of class, but the key is that they need to be graded. Without this accountability in the low-structure terms, students were not doing the reading and were likely cramming the week before the exam instead of distributing their study time. The second critical element seems to be encouraging the students in the class to view themselves as a community through small-group work in class. Further research could explore how best to approach in-class work to develop this sense of community rather than competition.

**Changes in Achievement, Behaviors, and Perceptions Vary among Student Populations**

In addition to looking at overall patterns in student behaviors and perceptions, we can also disaggregate these data to begin to understand why some groups might benefit more from the intervention. From the achievement data, we identified black and first-generation students as populations who responded most strongly to the treatment. Patterns in behaviors and attitudes were apparent for one of these populations (black students) and not the other (first-generation students).

The response of black students on our survey questions differed from other students in the class in three ways. First, under both classroom structures, black students were more likely to report that the homework contributed to their learning in the course, and there was a trend for black students to value the class to require more higher-order skills. These results imply that these students had a greater need for the kind of guidance provided by instructor-designed assignments. Thus, the addition of more homework and more explicit practice may have had a disproportionate impact on these students’ achievement. Third, black students were significantly less likely than other students to speak up in class, but this disparity disappeared under moderate structure. We suspect that the increased sense of the classroom as a community may have contributed to this increased participation.

Although first-generation students did not differ in how they responded to survey questions versus continuing-generation students, they could still differ in how valuable the changes in the course were to them. In particular, the increased sense of community that seemed to correlate with the implementation of moderate structure could have helped them disproportionately, as has been demonstrated in a previous study (Stephens et al., 2012). In addition, although students grouped in the category first generation share some characteristics, they are also very different from one another in terms of culture, background, and the barriers they face in the classroom (Orbe, 2004; Prospero et al., 2012). For example, in our university setting, 55% of first-generation students have parents with low socioeconomic status and 50% transfer in from community colleges. The variation in students could thus obscure any patterns in their responses. Future analyses will attempt to distinguish subpopulations to identify patterns potentially hidden in our analysis.

**Limitations of This Work**

One of the major purposes of this article is to recognize that classroom interventions that work in one classroom may not work in others because 1) student populations differ in how they respond to classroom treatments, and 2) instructors do not always implement the critical elements of an active-learning intervention. Thus, it is important for us to note that, although we have shown that increased structure can work with both majors and nonmajors and with students from a
range of racial and ethnic groups, we are still working in an R1 setting. More work needs to be done to establish the effectiveness of the increased course structure intervention in community college or comprehensive university settings (although the evidence that it works well for first-generation students is a good sign that it could transfer). In addition, this study was with one instructor, thus we can now say increased course structure has worked for two independent instructors (the instructor of the current course and the instructor of the original course; Freeman et al., 2011), but further work is necessary to establish its general transferability.

In addition, this study has suggested two factors by which increased course structure seems to be working by 1) encouraging distributed practice with a focus on class preparation and 2) helping students view the class as more of a community. Yet these are only two of many possible hypotheses for how this intervention works. It is possible that assigned preparatory assignments and small-group work to encourage community are not the only elements critical for this intervention’s success. Further studies could explore how to best implement activities in class or the impact of adding graded review assignments on achievement.

Implications for Instructor and Researcher Best Practices

As a result of implementing an increased course structure and examining student achievement and survey results, we identified the following elements critical for student success and the success of future implementations:

1. Students are not a monolithic group. This result is not surprising. Students vary in many ways, but currently we do not know much about the impact of these differences on their experience with and approach to a college-level course. Future studies on student learning should disaggregate the students involved in the study (if possible), so instructors looking to implement an intervention can determine whether, and potentially how well, a particular intervention will work for their population of students.

2. Accountability is essential for changing student behaviors and possibly grades. We found that without accountability, students were not reading or spending many hours each week on the course. With weekly graded preparatory homework, students increased the frequency of both behaviors. We did not provide them credit for reviewing each week, and we found the overall frequency of this behavior decreased (even though our results demonstrate that students who did review notes performed better).

3. Survey questions are a useful method of identifying what behaviors an instructor might target to increase student performance. From our survey results, it seems that creating weekly review assignments might increase the frequency that students review their notes and thus increase their grades. Without the survey, we would not have known which behaviors to target.

Overall, this work has contributed to our understanding of who is most impacted by a classroom intervention and how those impacts are achieved. By looking at the achievement of particular populations, we can begin to change our teaching methods to accommodate diverse students and possibly increase the effectiveness of active-learning interventions.

ACKNOWLEDGMENTS

This work was supported by NSF DUE 1118890 and DUE 0942215 and a Large Course Redesign Grant/lenovo Instructional Innovation Grant from the Center for Faculty Excellence at University of North Carolina (UNC). This study was conducted under the guidelines of UNC IRB Study 11-1752. We acknowledge support from the Center for Faculty Excellence at UNC, specifically Erika Bagley and Andrea Reubens, for getting the project off the ground with initial statistical analysis, and Bob Henshaw for his help with the collection of survey data. We also thank Mercedes Converse, Alison Crowe, Scott Freeman, Chris Lenn, Kathrin Stanger-Hall, and Mary Pat Wenderoth for their thoughtful contributions to early versions of the manuscript.

REFERENCES


National Science Foundation (2013). Widening implementation and demonstration of evidence-based reforms (WIDER). Division of Undergraduate Education NSF, 13–552.


Prospero M, Russell AC, Vohra-Gupta S (2012). Effects of motivation on educational attainment: ethnic and developmental...


Article

What Do I Want to Be with My PhD? The Roles of Personal Values and Structural Dynamics in Shaping the Career Interests of Recent Biomedical Science PhD Graduates

Kenneth D. Gibbs, Jr., †‡ and Kimberly A. Griffin †§

*Directorate for Education and Human Resources, Division of Human Resouces Development, National Science Foundation, Arlington, VA 22230; §Counseling, Higher Education and Special Education, University of Maryland, College Park, College Park, MD 20742

Submitted February 4, 2013; Revised June 26, 2013; Accepted June 26, 2013

Monitoring Editor: Vivian Siegel

Interest in faculty careers decreases as graduate training progresses; however, the process underlying career-interest formation remains poorly defined. To better understand this process and whether/how it differs across social identity (i.e., race/ethnicity, gender), we conducted focus groups with 38 biomedical scientists who received PhDs between 2006 and 2011, including 23 women and 18 individuals from underrepresented minority (URM) backgrounds. Objective performance and quality of advisor relationships were not significantly different between scientists with high versus low interest in faculty careers. Career interests were fluid and formed in environments that generally lacked structured career development. Vicarious learning shaped similar outcome expectations about academic careers for all scientists; however, women and URMs recounted additional, distinct experiences and expectations. Scientists pursuing faculty careers described personal values, which differed by social identity, as their primary driver. For scientists with low interest in faculty careers, a combination of values, shared across social identity, and structural dynamics of the biomedical workforce (e.g., job market, grant funding, postdoc pay, etc.) played determinative roles. These findings illuminate the complexity of career choice and suggest attracting the best, most diverse academic workforce requires institutional leaders and policy makers go beyond developing individual skill, attending to individuals’ values and promoting institutional and systemic reforms.

INTRODUCTION

The biomedical sciences (BMS) career landscape has transformed in the past half century. Forty years ago, the majority of PhD scientists progressed from graduate school to a faculty position, whereas today, only 14% of life sciences PhDs hold tenure/tenure-track faculty positions 5–6 yr after graduation (Stephan, 2012). Further, those pursuing an academic career face a significantly elongated path to independence—the average age at which new PhD investigators receive their first National Institutes of Health (NIH) RO1 grant is 42—and an academic job market in which the number of available tenure-track positions has not kept pace with the significant increases in the supply of newly minted PhDs (National Research Council [NRC], 2005b; McCook, 2011; Stephan, 2012). In line with these systemic changes, recent reports indicate many graduate students’ interests in research careers decreases as their training progresses (Fuhrmann et al., 2011; Sauermann and Roach, 2012), with many newly trained PhD scientists pursuing careers in policy, communication, law, and other nonacademic fields (Austin and Alberts, 2012).

Notwithstanding these changes, faculty members continue to play vital and unique roles in the scientific enterprise, shaping the national research agenda and training the next generation of scientists (Leggon, 2010). As such, funding agencies have invested significant resources in efforts to diversify the professoriate and research workforce (Handelsman et al.,
1. What is the process of career-interest formation toward or away from faculty careers? This study addresses two research questions: career decision making of 38 diverse recent BMS PhD graduates, with a particularly positive influence on the retention and persistence of students from underrepresented backgrounds (Hernandez, 2000; Matton et al., 2000; Jackson et al., 2003; Carrell et al., 2009).

Despite these efforts, diversifying the professorate remains "perhaps the least successful of the diversity initiatives" (National Academy of Sciences [NAS], 2011, p. 47), and the representation of women and minorities significantly lags in comparison with their share of PhDs awarded. African Americans, Latinos, Native American/Alaska Natives, and Native Hawaiian/Pacific Islanders (collectively underrepresented minorities, or URMs) receive nearly 13% of BMS PhDs awarded to U.S. citizens, but receive fewer than 5% of all NIH RO1 grants and are <2% of the basic science tenured tenure-track (TTT) faculty at U.S. medical schools (Nelson and Brammer, 2010; NRC, 2011; NIH, 2012b). Similarly, women receive more than 50% of life sciences PhDs, but their representation on basic science TTT faculty at the most research-intensive institutions remains around 33% (Jolliff et al., 2012). Thus, the diversity in the BMS PhD recipient pool is not reflected at the faculty level.

Strengthening the research enterprise and professoriate by ensuring the best researchers from all backgrounds participate requires a better understanding of the career decision-making process of recent PhDs, and how it might differ for scientists from underrepresented groups. Institutional hiring practices and climates have been cited as barriers to diversifying the faculty (NAS, 2011), and very recent evidence suggests that evaluation biases based on race/ethnicity and gender still exist, even when controlling for training and prior productivity (Ginther et al., 2011; Moss-Racusin et al., 2012). However, individual choice may also contribute to some of the underrepresentation seen at the faculty level. That is, highly skilled biomedical scientists from all backgrounds, including those underrepresented in faculty positions, may be choosing other career pathways because academic careers may be viewed as less attractive, welcoming, or satisfying than other professional options (NIH, 2012a).

This study begins to address the complex issue of career choice by examining the process and factors influencing the career decision-making of 38 diverse recent BMS PhD graduates, focusing specifically on their interest in becoming faculty. This study addresses two research questions:

1. What is the process of career-interest formation toward or away from faculty careers?
2. To what extent does this process differ based on social identity, specifically, race/ethnicity and gender?

THEORETICAL FRAMEWORK

Vocational psychology has developed and empirically validated theoretical frameworks to understand, estimate, and explain differences in career choice (Fouad, 2007), and these can be useful in understanding career decision making in BMS. The social cognitive career theory (SCCT), and the social influence model frame this study. SCCT has been used to explain significant amounts of variance in the career choices of undergraduates from majority and URMs backgrounds in science and engineering education (Lent et al., 2005, 2008, 2011; Byars-Winston et al., 2010) and is the basis of current interventions designed to broaden participation in the professoriate (Byars-Winston et al., 2011).

SCCT frames career attainment as a developmental process in which individuals make a series of personal decisions shaped by social and institutional context (Lent et al., 1994). SCCT posits that interests (i.e., "Do I want to do this?") lead individuals to pursue a particular training path (choice goals) and then to undertake the courses of action necessary to attain that goal (choice actions). Durable career interests are thought to result from positive self-efficacy beliefs and outcome expectations. Self-efficacy describes an individual's belief in his or her capacity to successfully execute the courses of action needed to achieve a goal ("Can I do this?"), while outcome expectations frame an individual's anticipated outcomes for pursuing certain courses of actions ("What will happen if I do this?"). SCCT also recognizes the roles that personal characteristics (e.g., gender, race/ethnicity, disability status), learning experiences (e.g., access to role models, faculty or peer discouragement), and contextual supports and barriers can play in the process of career choice and development.

This study examines whether and how recent PhD graduates describe the roles of self-efficacy, outcome expectations, and learning experiences in shaping goals and courses of action toward or away from pursuing faculty careers, as well as whether and how these vary based on social identity.

Recent work has also used a social influence model to frame the career intentions of undergraduate and graduate science students from URM backgrounds (Estrada et al., 2011). Estrada and colleagues showed that self-efficacy became a poor predictor of intentions to pursue a scientific career when "identification as a scientist" and "internalization of [scientific] values" were also considered. In this context, values refer to what matters and is important to the decision maker, and identity refers to the extent to which the decision maker has adopted the roles of and feels connected to a profession. In the face of challenges and barriers, it is important to consider the extent to which scientists from all backgrounds, including those underrepresented in the professoriate, make decisions about their interest in research and academia based on their values and the extent to which their values align with the values of academic culture. Thus, this study also considers whether and how individual values and the perceived values of the academy may play a role in the articulation of career goals, choice actions, and career decision making.

METHODS

Data Collection and Procedures

The purpose of this study was to examine how scientists who completed BMS PhDs between 2006 and 2011 made decisions about their career paths and to explore the extent to which decision making varied by social identity group. Therefore, a purposeful sampling strategy (Merriam, 1998) was developed to recruit a diverse set of participants with respect to
both social identity (i.e., race and gender) and career interest (i.e., faculty and nonacademic). Participants were recruited through listservs of PhD-level, science-policy professionals and academic and government postdocs; via direct contact at national scientific conferences; and with the help of graduate student and postdoctoral administrators at three research universities. Participants were also asked to recruit other eligible peers (i.e., snowball sampling; Bogdan and Biklen, 2007).

Potential participants completed a brief, online questionnaire, reporting demographic and academic information. Participants were also asked to indicate their interest in each of the following career tracks on a 5-point Likert scale: academic (research intensive), academic (teaching intensive), research career (nonacademic), or nonresearch career (e.g., consulting, policy, science writing, etc.). On the scale, 1 was equivalent to low interest, 3 represented moderate interest, and 5 indicated high interest. Thus, participants who answered 4 or 5 on either academic option (research or teaching intensive) were classified as having “high” interest in faculty careers, while those who answered 1–3 for both types of academic positions were classified as having “low” interest.

Eleven focus groups were conducted. Focus groups are ideal for promoting deeper understanding of how people think and feel about issues, experiences, and ideas (Kruger and Casey, 2000). Efforts were made to match individuals with similar levels of interest in faculty careers and, when possible, based on membership in a URM group. Groups ranged from two to six participants and met for 60–90 min. Participants were engaged around a series of questions exploring the development of their interest in science, changes in their career aspirations, graduate school and postdoctoral training experiences, and current professional goals. The protocol was semistructured, fostering consistency across focus groups but also allowing researchers to explore unique emerging phenomena. With participants’ permission, all focus-group interviews were audio-recorded and transcribed. All participants were assigned and are referred to by pseudonyms. All work was done in compliance with the Pennsylvania State University (PSU) Office of Research Protections IRB #38896 (K.A.G. was affiliated at PSU during the period in which this research was conducted).

Data Analysis
Demographic, academic, and career-interest data were used to develop a profile of each participant and were analyzed by comparing the level of interest in academia based on objective performance measures, such as time to degree, number of publications, and training institution. Statistical analyses ($t$ tests) were performed in SPSS (www.ibm.com/software/analytics/spss/).

The analytical design was informed by the team-based strategy used by researchers at the Centers for Disease Control (MacQueen et al., 1998), with the goal of generating themes, trends, and understandings of common experiences across participants in the study (Bogdan and Biklen, 2007). Qualitative data were organized through a systematic, multistage process, consistent with the constant comparative method (Glaser and Strauss, 1967). Preliminary memos were written after each focus group to highlight first impressions of emerging themes. After interviews were transcribed, the transcripts and preliminary memos were read closely, and another set of analytical memos was written based on each focus group. Then comprehensive memos were drafted, identifying themes across all of the focus groups. Comprehensive memos were compared and discussed, and were then combined into an initial list of phenomena emerging from the data, capturing the ways in which PhD scientists described their career decision making. An initial list of inductive codes was developed based on these themes. The initial list of codes was supplemented through a deductive process, wherein codes capturing interview questions, propositions of SCCT, and the literatures on the development of faculty aspirations and academic socialization were added. This assisted in the process of identifying disconfirming evidence (inconsistent with emerging themes and overarching trends in the data), capturing the full range of participants’ responses to questions.

Transcripts were then reread and coded. Code reports were generated, allowing the researchers to examine data assigned the same code and compare quotations with early perceptions of qualitative themes emerging from the data. To ensure trustworthiness of findings, the researchers were mindful of recognizing discrepant or disconfirming evidence, or examples not consistent with presented explanations and interpretations (Maxwell, 2005). Thus, in addition to presenting evidence supporting a theme, the researchers sought data inconsistent with the themes when establishing the salience and consistency the findings. Qualitative themes with similar underlying principles were clustered together, reflecting the ways in which various factors and experiences shape scientists’ interest in pursuing careers as faculty. Larger narratives were then developed about the nature of participants’ experiences during each training period and at each decision point, using data to support assertions.

RESULTS

Study Participants
Demographic and training information about study participants is found in Table 1 (full information on participants is found in Supplemental Table S1). The cohort included 38 scientists who received their PhDs at 28 different institutions between 2006 and 2011: 36 in the basic BMS, one in behavioral sciences, and one in other life sciences (NIH, 2012a). Eighteen participants (47%) belonged to URM groups (14 African American/black, 4 Hispanic/Latino), while 20 belonged to groups well represented in the BMS (17 white, 12b 15

<table>
<thead>
<tr>
<th>Interest in faculty position</th>
<th>High ($n = 19$)</th>
<th>Low ($n = 18$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>14</td>
<td>9</td>
</tr>
<tr>
<td>Male</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>URM</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>Received PhD from top-50 funded institution</td>
<td>12b</td>
<td>15</td>
</tr>
<tr>
<td>Completed postdoctoral training</td>
<td>16</td>
<td>11</td>
</tr>
<tr>
<td>Currently a postdoctoral trainee</td>
<td>13</td>
<td>6</td>
</tr>
</tbody>
</table>

*One participant did not complete the demographic questionnaire but self-identified to the investigators as belonging to a URM group.

*One participant completed the PhD at an overseas institution.
3 Asian). Twenty-three were women (11 from URM backgrounds and 12 from white/Asian backgrounds), and 15 were men (7 from URM backgrounds and 8 white). Twenty-eight completed postdoctoral training, including 20 who were postdocs at the time of the study. All participants were U.S. citizens/permanent residents, and all except one (Judy) completed graduate training in the United States. Of those completing PhD training in the United States, the majority (73%) received PhDs from one of the top-50 funded research universities.1

**Quantitative Analysis**

**Objective Performance Measures Did Not Explain Differences in Interest in Faculty Careers.** In this sample, the research productivity of the scientists with “high” versus “low” interest in faculty careers was comparable (Table 2). Scientists with high interest in faculty careers had a median of two first-author and seven total peer-reviewed publications (indexed in Scopus2); the journals in which they published had an average impact factor of 5.8; and these scientists had a mean h-index of 5 (Hirsch, 2005). Scientists in this sample with low interest in faculty careers had a median of 2.5 first-author and six total publications; the journals they published in had an average impact factor of 6.1; and these scientists had a mean h-index of 4.7. None of these differences were statistically significant (p > 0.05, two-tailed t test). Neither were differences in median time to degree significantly different between those with high versus low interest in faculty careers (6 vs. 5.25 yr, respectively). Therefore, for this sample of scientists, traditional objective performance measures did not account for reported differences in interest in faculty careers.

**Relationship with Research Advisors Did Not Explain Differences in Interest in Faculty Careers.** The nature of these scientists’ relationships with graduate advisors also did not directly relate to their interest in faculty careers. The percentage of participants using positive language to describe their advisor relationships (i.e., terms such as “amazing,” “very supportive,” “great”) versus those using mixed or negative language to describe their relationships (e.g., terms such as “rocky,” “friction,” “difficult”) was not significantly different for those with high versus low interest in faculty careers (Table 3). Of the participants with a high interest in faculty careers, 42% described their relationships with their advisors positively, 47% described them negatively, and 11% did not describe the nature of their relationships with their advisors in great depth. In comparison, 56% of those with low interest in faculty careers described their relationships with their advisors positively, while 28% described their relationships negatively, and 16% did not describe the nature of their relationships with their advisors in great depth. Differences between two groups were not statistically significant (p > 0.05, two-tailed t-test). Thus, differences in the nature of their relationship with their PhD advisors also did not explain differences in interest in pursuing academic careers.

**Narrative Analysis**

Narratives from the 38 participants were examined to understand the process of career-interest formation. A thematic analysis, capturing trends across participants in the study, is presented below, describing: 1) choice actions with respect to training decisions; 2) training environments; 3) learning experiences; and 4) career choice goals, especially the decisions these scientists made toward or away from faculty careers. Unless otherwise noted, the themes described below were consistent across scientists from all racial/ethnic and gender backgrounds.

**Choice Actions with Respect to Training Decisions Are Often Not Linked to Clear Career Goals**

Pursuit of a PhD Is Often Motivated by a Love of Science and Career Potential, Not a Specific Career Goal. For most participants in this sample, the decision to pursue a PhD was not linked to a specific, well-defined career goal (Figure 1A): 55.2% described entering their PhD training with undefined (i.e., not clearly defined) career goals, while 26.3% stated they entered with intentions to pursue a faculty career (although half of this group said their interest in a faculty career was primarily shaped by a lack of knowledge of the other career options for PhD scientists). Remaining participants indicated they planned to pursue careers in academia or industry (10.5%), or a research career outside academia (i.e., industry or government; 5.3%). One scientist did not describe her career interests at entry.

Two core reasons were consistently articulated for the decision to pursue a PhD: 1) a love of science, typically fostered by undergraduate research; and 2) the recognition that a PhD was required for career mobility. Darren, a postdoctoral scientist, noted both of these reasons in his decision to pursue a PhD:

---

Table 2. Objective performance measures for study participants

<table>
<thead>
<tr>
<th>Interest in faculty position</th>
<th>High (n = 19)</th>
<th>Low (n = 18)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median first author publications (range)</td>
<td>2 (0–12)</td>
<td>2.5 (0–7)</td>
</tr>
<tr>
<td>Median total publications (range)</td>
<td>7 (2–26)</td>
<td>6 (1–19)</td>
</tr>
<tr>
<td>Mean impact factor, journal publications (range)</td>
<td>5.8</td>
<td>6.1</td>
</tr>
<tr>
<td>Mean h-index (range)</td>
<td>(0.76–36.28)</td>
<td>(0.73–32.4)</td>
</tr>
<tr>
<td>Median years to PhD completion (range)</td>
<td>5 (1–11)</td>
<td>4.7 (1–10)</td>
</tr>
<tr>
<td>Median first author publications (range)</td>
<td>6 (4–8)</td>
<td>5.25 (4–7)</td>
</tr>
</tbody>
</table>

Table 3. Participants’ descriptions of nature of relationship with PhD advisor

<table>
<thead>
<tr>
<th>Interest in faculty position</th>
<th>High (n = 19)</th>
<th>Low (n = 18)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>42%</td>
<td>56%</td>
</tr>
<tr>
<td>Mixed/negative</td>
<td>47%</td>
<td>28%</td>
</tr>
<tr>
<td>Not described</td>
<td>11%</td>
<td>16%</td>
</tr>
</tbody>
</table>


2For more information on the Scopus search engine, see: www.info.sciverse.com/scopus.
Training decisions often not linked to clearly defined career goals. Pie chart showing the top codes emerging from analysis of participants' narratives with respect to their career goals at entry into (A) PhD training or (B) postdoctoral training. Scientists who indicated that they did not have a clearly defined career goal are labeled as “undefined” and shown in blue; those who indicated interest in faculty careers are shown in red; those who indicated interest in either an academic or industry career are shown in green; and those who indicated they had an interest in a research career outside of academia (e.g., industry, government, etc.) are shown in purple. Participants who did not clearly describe their career interests are shown in gray. Ten scientists in the sample did not complete postdoctoral training and are not included in (B).

Figure 1. Training decisions often not linked to clearly defined career goals. Pie chart showing the top codes emerging from analysis of participants’ narratives with respect to their career goals at entry into (A) PhD training or (B) postdoctoral training. Scientists who indicated that they did not have a clearly defined career goal are labeled as “undefined” and shown in blue; those who indicated interest in faculty careers are shown in red; those who indicated interest in either an academic or industry career are shown in green; and those who indicated they had an interest in a research career outside of academia (e.g., industry, government, etc.) are shown in purple. Participants who did not clearly describe their career interests are shown in gray. Ten scientists in the sample did not complete postdoctoral training and are not included in (B).

I did undergraduate research my entire four years [and] never really wavered on going to do a PhD. Part of it was, I saw that the people who moved up in science ... whether they were in academia or private industry, all had a PhD at the higher levels. So instinctively, I felt that if I wanted to be at one of the higher levels, then at some point I would need to get a PhD. But also, I just really loved working at the lab bench, the discoveries ... I just really enjoyed working at the lab bench.

Leticia similarly pursued a PhD, because she “fell in love with research as an undergraduate.” Meanwhile Matt’s undergraduate research experiences showed him “that if you wanted to do anything in science, you really needed to get your PhD—to have credibility and to do your own research.” Collectively, the narratives showed the main drivers of PhD attainment were the affinity toward research and the outcome expectation that a PhD was required for professional mobility and independence.

Participants differentiated between their love of research and intending to pursue a particular career path—faculty or otherwise. Meredith describes her mind-set at the beginning of her graduate studies this way:

I did not have an end goal in mind ... I don’t think I ruled out being an academic research professor. But it wasn’t the only reason I was going ... I really enjoyed biology, I enjoyed asking questions, I enjoyed the experimentation process ... but I didn’t really have an idea of what I wanted to do in the end.

Similarly, Jordan began his PhD without “a clear end point” with respect to his career goals, but perceived the PhD “would be a springboard to new career possibilities,” while Robin said, “I never thought about what I would do with a PhD ... until right near the end [of graduate school].” This lack of career certainty was also true for the two men in the sample who had sought and recently accepted faculty positions at the time of the focus group. Mark said he “didn’t really have any [career] goals” going into graduate school, while Thomas thought he would “get a PhD and then find a job” in industry, noting specifically that he “wasn’t thinking of an academic professorship at that time that I entered the PhD program.” Thus, participants chose to attend graduate school based on a love of science and an interest in scientific career mobility rather than a commitment to a specific career, faculty or otherwise. Moreover, very few participants articulated a strong, early commitment to pursuing faculty careers.

Pursuit of Postdoctoral Training Is Often Independent of a Clear Career Goal. As was the case with earning a PhD, participants distinguished between pursuing postdoctoral training and intending to pursue a faculty career. Again, the majority of focus-group participants (57.1%) described entering postdoctoral training without a clearly defined career goal (Figure 1B). For some, postdoctoral training provided an opportunity to clarify career plans. Nancy was unsure of her career goals after graduate school and saw postdoctoral training “was a great way to keep all of my options open.” Meanwhile, Melanie chose to postdoc, because “I didn’t know what I was going to do when I left grad school. I just knew I didn’t want to stay in academia.” While she still enjoyed bench science, the postdoc allowed her time to “figure out exactly what I’m going to do.”

For many of the scientists who began postdoctoral training without a clearly defined career goal, the postdoc represented a default pathway rather than an intentional choice, made without strong consideration of other options. For example, Eric said he “didn’t really consider anything other than a postdoc, but then may be naïve so,” and Claudia seemed to default into postdoctoral training, asking rhetorically, “What was I gonna do when I finished my PhD other than do a postdoc?” Still others chose to pursue academic postdocs because of their belief that any other career step would foreclose future opportunities in academia. While Ann was unsure at the end of graduate studies that she wanted to be a faculty member, she chose to do a postdoc because of her perception that “if you branch out into industry or writing or science policy, you can’t go backward into academia.” Robert shared similar thoughts, noting that, while other career options were of interest to him, “I started having second thoughts about how reputable a postdoc in industry would be later if I chose to go back into academia.”

While 17.9% pursued postdocs because of their goal of obtaining a faculty position, 10.7% used the postdoc as a way to become more competitive for research careers outside academia, that is, in industry or government (Figure 1B). Rodney decided in graduate school not to pursue a faculty career and chose to do a postdoc “to finish some of the papers” and get a letter of recommendation, so he could “then go do something else.” Meanwhile, Natalie found that most of the government research positions she wanted required 3–5 yr of research experience beyond the PhD, so she decided to “do a couple postdocs” to “beef up” her skill sets for the research careers outside academia that she desired. Collectively, the narratives indicate that postdoctoral training opportunities were most often seen as a way to keep one’s career options...
open, rather than necessarily indicative of a particular career choice, or a declaration of one’s intent to pursue an academic position.

Training Environments Generally Lacked Structured Career Development. While many participants entered their PhD and postdoctoral training with malleable career interests, only 16% of focus-group participants described having departmentally or institutionally sponsored, structured career development activities throughout their training—either toward or away from academic positions. Alicia, a postdoc with high interest in a faculty career, noted she did not learn “about what’s being called ‘alternative career options’” until reaching her postdoctoral institution. Similarly, Jack, a research scientist in a federal laboratory, noted that in graduate school “there weren’t a lot of resources around to learn about what [career] possibilities were out there.”

Poorly structured career development was especially pronounced for the academic postdocs in the sample. The NIH and NSF define postdoctoral positions as “temporary and defined periods of mentored advanced training to enhance the professional skills and research independence” for doctoral degree holders “to pursue [their] chosen career path” (Bravo and Olsen, 2007). Most postdocs in this sample, particularly those in the academic arena, described their experiences as lacking structured professional development. As Erin, a current postdoc, said,

I am in a training position, and told that I am a “trainee,” that’s on my title. I can’t think of a training moment that I have been a part of, where I was being trained.

Similarly, as Judy reflected on her postdoctoral experience, she rhetorically asked,

If we are trainees, shouldn’t there be some sort of structure for us to help guide us or teach us what it is that we’re meant to be learning as trainees? It really baffled me. We get nothing.

Daniel added that, from his point of view, “being in the lab alone doesn’t train us to be good scientists. We need other training beside technical work to become good scientists.” All three of these postdocs were interested in pursuing faculty careers but felt they needed to develop additional skills such as grant writing, teaching experience, budget management, and personnel management before transitioning to a faculty position. Independent of their career interests, the academic postdocs in this sample described a lack of structured opportunities to develop these skills as part of their training, consistent with previous reports critiquing the lack of structured training postdocs receive (NAS, 2000; Davis, 2005). The notable exception was Thomas, a postdoctoral scientist at the NIH, who pointed out the presence of “an entire office … devoted to helping people make career decisions” at the NIH, but observed that, in his experience, this support “certainly is lacking at other institutes.”

Learning Experiences Shape Outcome Expectations, and Differ by Social Identity

Vicarious Learning Shapes Outcome Expectations about Academic Hiring. In a context in which there was little structured career development, vicarious learning experiences (i.e., learning by observation) played a strong role in shaping outcome expectations about the process of obtaining a faculty position. Scientists in this sample—indeed of career interest—articulated a nearly uniform outcome expectation that obtaining a faculty position was extremely difficult and often not predicated principally on research productivity. Thomas described obtaining his faculty position after 6 yr of postdoctoral training as the result of luck and chance rather than the realization of hard work:

I just kept at it hoping that something would eventually come my way. I didn’t delude myself at the beginning of this … the odds are greatly stacked against me. … I am just going to keep playing the lottery and maybe it will come my way. And one eventually did.

Similarly, Eric, a first year postdoc, aspired to a faculty career, despite perceiving it to be an increasingly unlikely proposition:

The longer I have been in this, the more I realize that not everyone that’s good and works hard is able to make that transition from postdoc to faculty. I guess that naively, I assumed if you worked hard and did good work, you would make it. Naively, I guess I still believe that, because I have to, but it’s becoming increasingly clear that’s not true for everybody.

Audrey, who also had a high interest in a faculty position, echoed the sentiments that she perceived obtaining a faculty job as not being predicated principally on research accomplishments. Having been a member of faculty search committees throughout her training, she remarked that, based on her observations, “it just seemed like at the end of the day” getting a faculty job “is not really necessarily based on merits, but who you know and how well you can network and play with the system.”

Those with low interest in faculty careers echoed similar sentiments, noting observations of the extreme difficulty in obtaining faculty positions in the current landscape. Will described the employment prospects of a postdoc from his PhD thesis lab who he perceived as highly talented:

I had a postdoc in my lab who had three Nature publications. Three! It took him two-and-a-half years to find a job. He is a very smart person; very good with mentoring … But I am like, if that type of person, [who] did this well cannot find a position … He ended up going to [an institution] and it is obvious that he didn’t want to live there, but he had been a post doc for six years. He just had to move on. He had no control whatsoever in that.

Others with low interest in faculty careers pointed to the low supply of faculty jobs relative to the number of trained scientists as a source of difficulty in obtaining faculty positions. Melanie said, “There’s obviously not enough academic jobs for all these grad students [and] all these postdocs that are coming through.” These sentiments were echoed by Matt, who said, “There are not enough faculty jobs for people who want to stay in academia.” Thus, the scientists in this sample—across career interests and social identity—described an outcome expectation that obtaining an academic position in the current job market is extremely difficult and often not predicated principally on performance, with very talented scientists having extreme difficulty obtaining faculty positions.
Vicarious Learning about the Nature of Faculty Lifestyles Shapes Expectations about Professional Outcomes. Independent of their level of interest in a faculty career, most participants described how their observations of professors at major research institutions created an outcome expectation that faculty lifestyles at major research universities were extremely demanding. Judy, who has a high interest in a faculty position, described her postdoc advisor’s life as one in which he is “traveling around the world, not seeing his family, and sending out emails at three or four in the morning.” Alicia, who also has high interest in a faculty career, noted the demanding workload of faculty, saying, “It’s very intimidating to think about the amount of time that I will have to spend to make sure my lab stays up and running, and that I don’t have grad students or postdocs who are . . . worried about not having the funding to do the work and to live, to eat.”

Those with low interest in faculty careers also shared these sentiments. Aaron described the schedules and lifestyles of faculty at his institution with whom he rotated. One was “HHMI funded and pumping out Cell [and] Nature papers,” but the demands on his time meant “he was never around,” while the other was a “young tenure-track professor” with “really sexy projects,” but his faculty obligations meant he had no “outside life.” Similarly, Mandy, who has a low interest in a faculty career, noted that in observing the lifestyles of her principal investigator (PI) and other faculty in her graduate institution, “They are guiding other researchers, writing grants, and going home very late at night, every night. So it was like, ‘if that’s what I want to do, these are the people who I will turn out like. These are the lifestyles that I’ll have to follow if I want to stay in research.’” Thus, across career interests and social identity, and consistent with previous work (Mason et al., 2009; Fuhrmann et al., 2011), participants noted that, based on their vicarious learning experiences, faculty life was demanding and required extensive commitments in terms of time and energy. Notably, only two participants (5% of the total sample) described faculty role models at major research universities who lived balanced lifestyles.

Unique Learning Experiences for Women and Scientists from URM Backgrounds. In addition to these shared learning experiences and outcome expectations, scientists from groups underrepresented in the professoriate (women and URM) described unique learning experiences and outcome expectations shaped by their race and gender. Three-quarters of the female scientists spoke about gender directly shaping either their training experiences or outcome expectations about faculty life, including all of the women with a high interest in faculty careers. Nancy recounted her experience at conferences in her male-dominated subdiscipline, where she, as one of the few women present, would receive “invitations back to faculty members’ hotel rooms” after the networking activities. This in turn began to impact her self-efficacy, as she wondered, “Was this faculty member talking to me all evening because he’s interested in my science or because I’m the only woman in the room and he wanted to go to bed with somebody?” Similarly, Erin described experiencing “major sexual harassment” from her PI as a graduate student, which included the PI adding a male author on her paper, because, in his words, he could not “listen to what a female was saying and believe it.” Moreover, Alicia, a postdoc and mother who trained in an engineering-related subdiscipline, described the strong “pressure against having children” she had experienced throughout her training, and an instance in which she had to confront a colleague who wanted to hire a male undergraduate student instead of an equally qualified female student, because, in his view, “women can’t do math and they’re not really competent in mathematics.” These and other women described learning experiences directly shaped by gender in a manner that the men in the sample did not describe.

In addition to these direct experiences, vicarious learning experiences of female professors also shaped the outcome expectation that pursuing a faculty career as a woman presented additional challenges. Christina noted that it seemed that “in order to be taken seriously you have to either overly assert yourself, or suppress your femininity.” However, Nancy had a different perspective and felt there was a “double standard” when women assert themselves as compared with men, a view that was shaped by her observation of two “aggressive young faculty”—one male and one female. Although, she said, both “tended to make their students cry,” the female professor was characterized in the department as a “raving [expletive],” while the behavior of the male professor was excused. Claudia noted how the contributions of women were generally minimized when compared with their male colleagues. Male invited seminar speakers were called “experts” by department faculty, while the female speakers were called “bright,” a term she felt was more appropriate for a “high school student” than for well-funded and well-published faculty members. Thus, women in the sample—including all of the women with high interest in faculty careers (such as the five above)—described gender-specific training experiences or outcome expectations resulting from direct or vicarious learning, leading to expectations of encountering gender-related barriers that male colleagues would not if they pursued faculty careers.

There were also racial differences in training experiences, with the influence of race/ethnicity described in more personal and profound ways for women of color. For men of color, race was acknowledged as a reality, but they did not describe it as demonstrably impacting their training experiences. For example, Steve described his PhD cohort as a “rainbow coalition” and said he had “no race issues,” while Tim said that, with the exception of being mistaken for other minority scientists at conferences, race was not a major factor in his training and he was treated like “a regular grad student.” In contrast, women of color more often described race as having a direct impact on the quality of their training experiences. Robin said plainly, “Race is definitely an issue” in her graduate training. On returning from a departmental meeting, her PI once said to her, “It is really important to the department for you to get out. [The department] really wants minorities to get out [i.e., graduate] because of numbers.” Although she noted that he was simply relaying facts and was not speaking in a “derogatory” manner, it began to impact her self-efficacy. As she explained, “[it] definitely makes you second-guess [yourself] . . . it affects how you perceive how good you are, when someone tells you those sorts of things.” Race also shaped Natalie’s training; she noted that, as the “only person of color” in her postdoctoral department, she consciously went out of her way, possibly sometimes to her “detriment,” to ensure she was not portraying any
That Vary Based on Social Identity. Scientists in this study had Interest in Pursuing Faculty Careers Is Driven by Personal Values and Structural Dynamics. Career Choice Goals Influenced by Personal Values and Structural Dynamics. Interest in Pursuing Faculty Careers Is Driven by Personal Values That Vary Based on Social Identity. Scientists in this study had similar objective levels of performance, perceptions and outcome expectations about faculty careers, and limited exposure to structured career development. These factors did not appear to explain development of interest in faculty careers. Notably, the primary driver for pursuing a faculty career, across racial, ethnic, and gender backgrounds, was personal values. That is, if what a scientist felt was important to him or her personally and professionally could be best achieved in an academic setting, he or she pursued a faculty career. These scientists described pursuing a faculty career despite their recognition of the structural dynamics of the current biomedical workforce (e.g., a difficult academic job market, low grant-funding rates, long training periods, and extremely high faculty workload) or the challenges faced specifically for scientists from groups underrepresented in the professoriate (e.g., training experiences and outcome expectations adversely impacted by social identity).

The values motivating pursuit of a faculty position were multifaceted and varied largely, but not exclusively, by social identity. Some scientists, particularly those from racial/ethnic majority backgrounds, articulated their choice goals based on the perception that academic environments allowed the greatest freedom to pursue research topics of interest. For example, Thomas began his PhD with the choice goal of obtaining a position in industry, but was inspired to pursue a faculty career when he attended scientific conferences in graduate school and began to understand “what academic research and teaching really meant.” He had thought of professors as functioning in a manner similar to high school teachers, but was drawn to the professoriate when he saw that faculty members were doing “really mind blowing science” and by the “degree of freedom that came with academic research.” Similarly, Mark was drawn to the professoriate because of “the freedom of academic research… just getting to do whatever you want, and have people pay you… That is awesome!”

Other postdocs from majority racial/ethnic backgrounds also articulated research freedom as the reason they continued to pursue academic careers. Eric said he continued to be drawn to a faculty position because of “the science, the questions, [and] the flexibility” he perceives the career path offers. Similarly, Ann continued to pursue an academic career because of the freedom to pursue her own research ideas, saying, “If I could have my own lab with a small group of people and pursue the research that I think is interesting… that would be fabulous.” An industry career was not an appealing choice goal for her because of her perception that “they can take your projects away from you at any time.” Thus, for many scientists from racial/ethnic majority backgrounds, including all the men and one woman who had attained or were pursuing faculty careers, research freedom was articulated as the main motivator for choosing to pursue an academic career.

In contrast, nearly all other women and postdocs from URM backgrounds who had high interest in faculty positions were motivated by the ability of faculty to engage in externally focused values they deemed important. This went beyond an interest in basic science or the freedom of academic research described by their peers from well-represented racial/ethnic or gender backgrounds. For example, some were motivated by the value they placed on the application of research to health problems facing their communities. Christina worked in a research field whose findings directly impact clinical practice in women’s health, one in which, in her words, poorly designed studies can have “detrimental” consequences for patients and “failure actually means changing someone’s life.” While she described her field as “fascinating,” she was persisting on the academic track and maintaining her faculty choice goal because “I’m really interested in where we are [and] getting to the finish line. That’s why I come into work every day.” She noted that “the work is so much bigger than any of the individuals” she works with, which kept her from letting the “negative experiences” she has had with advisors and colleagues deter her from continuing in the field. Similarly, Daniel felt that “an academic setting” would offer him a better environment in which to explore the research questions he felt were necessary to address a health challenge that is particular to his ethnic community. Therefore, for some scientists from URM backgrounds, the perception that academia presented them with the most favorable environment to conduct research applicable to pertinent health problems in communities they wanted to serve drove them to continue pursuing a faculty position.

Others placed a high value on impacting students, and the roles faculty members play as mentors motivated their career choice goals. Robert said he sees “few people of color in academia” and felt pursuing a faculty job was “a responsibility… I can do this so I should.” Alicia had an academic career as her goal, because “It’s rewarding to think about being able to mentor students. That’s really what I’m passionate about.” Nancy, who had accepted a position at a liberal arts college, said she was “very motivated to make sure that undergrads have that exposure to real research,” noting “the biggest difference you can make” in someone’s life is to “go be a good mentor.” Similarly, Erin, a second-year postdoc who planned to pursue a faculty position at a research-intensive university was motivated by the fact that “research institutions need people that care about the way [training] is done… that is what keeps me going.” Claudia was continuing on the path to a faculty career despite a “very rocky” graduate school experience, because she “knew that if I stuck this out that I could create opportunities” for others in science. Thus, many of the scientists in this sample from groups underrepresented in the professoriate (women and URM) placed importance on the externally focused values of serving as role models and facilitating the success of students; the ability of faculty to play these roles strongly motivated their pursuit of faculty careers.

Nonacademic Career Choice Goals Are Driven by Values and Structural Dynamics. Scientists pursuing careers outside academia also cited personal values, as well as the structural dynamics faced by PhDs seeking faculty positions, as the
drivers of their career decision making. Scientists who felt that the nature of faculty work would not allow sufficient engagement with values they felt were important chose nonacademic career paths. The principal value articulated—across social identity—was the need for their work to have a higher level of applicability than they felt would be attainable in a university. Mandy, a woman from a majority racial/ethnic background, exemplified the sentiments of many of those moving away from academia, saying,

I wanted to make a difference in the world with my career, and I didn’t see the molecules that I was studying actually changing anyone’s life.

Similarly, Trent, a male from a majority racial/ethnic background, said that he began his PhD training “completely excited” about basic research but grew frustrated at the length of time needed to translate “basic research to any practical end result.” His choice goal was to have a career in science policy, because, in his view, “it’s an expression of science that engages directly into the public realm.” Jack, also a male from a majority racial/ethnic background who worked as a scientist in a federal research lab, said that, while in his view basic research is “great,” if his work did not lead to a “tangible advance, a product or a medicine or something like that,” then he did not see “the point.” He noted that in his current research position outside of academia, he had been able to pursue his goals of applying research to tangible outcomes “quite well,” noting that at the end of his projects “we actually have something you could touch.”

Several women of color also cited the need for their work to have more directly applicable outcomes as the reason for their diminished interest in faculty careers. Lara moved into policy after postdoctoral training, because of “a misalignment in terms of my goals, the goals of the field and the goals of my department.” While some saw academia as a means to help their communities in a direct way, she noted that “the reward structure of academia” would not allow her work to impact her own community in the way she felt it was important. For others, working in academia limited their ability to translate research findings into practical application, making it less appealing. Robin, also a woman of color, “bowed out” of pursuing a faculty position because she felt academic work “was so incrementally small compared with the big picture in terms of where I thought I wanted to make a change.” In sum, many of the scientists who moved away from academic career paths and developed other choice goals—dependent on the nature of faculty work would not allow them to make the types of contributions they deemed important.

In addition to personal values, the structural dynamics faced by scientists seeking academic employment also influenced choice goals, in some cases discouraging interest in faculty careers. For two-thirds of all men pursuing nonacademic careers, including all of the men from URM backgrounds, cited structural dynamics in the workforce as important reasons driving their career decisions. The structural dynamics motivating some women to choose nonacademic careers were distinctive from those described by men. Career–life balance issues loomed large, and for some were determinative in their decision to pursue nonacademic careers, consistent with previous reports (Macon et al., 2009; Fuhrmann et al., 2011). Melanie said, “I saw the amount of stress and amount of total devotion to your lab and your research and writing grants . . . the amount precedence that [a faculty career] took over every other aspect of your life” and decided she “couldn’t stay in academia.” Similarly, Mandy, in addition to the limited impact she felt she would have as a faculty member, perceived an academic lifestyle as incompatible with other priorities, such as having a family. She noted, “I want to have a family and to be able to see my children grow up, not just get home when they are in bed” but felt that would not be possible as a professor, solidifying her pursuit of a nonacademic career. For Deborah, a woman of color, the climate she faced during her training prompted her move away from the professoriate. She said her interactions with colleagues were “the thing that deterred me from seeking a career in an academic environment,” noting “there’s no reason for me to work really hard to have these people be my peers.” Thus, in addition to values, a number of structural dynamics in the research workforce and university environments prompted many scientists to pursue nonacademic careers.

DISCUSSION

While large numbers of scientists continue to pursue academic careers, the academy will continue to face significant challenges moving forward—especially in ensuring that the next generation of faculty includes talented scientists from a wide range of backgrounds. This study presents analysis of data documenting the process of career-interest formation and the factors influencing this process, collected from focus groups conducted with a diverse group of productive American PhD biomedical scientists. The data from this exploratory study are presented in an effort to better understand the process of career-interest formation and how this process differs by social identity, providing an emerging framework that can be tested in future work and against which others’ experiences can be examined and compared. This work is the first stage of a larger project, which will examine whether
findings presented in this article are consistent across a larger sample and explore emergent themes with more depth. Subsequent research integrates quantitative data collected from almost 1900 participants and interviews with a subset of a diverse sample of 125 PhD biomedical scientists, documenting their training experiences and career development processes.

Participants in this study described career-interest formation as a dynamic process. Previous research and efforts to broaden participation in the nation’s academic workforce have highlighted the importance of faculty relationships and mentorship, self-efficacy, and research training experiences (Lindley, 2006; Burke, 2007; Kaiser, 2012a; Wadman, 2012). However, neither these constructs, nor objective performance measures traditionally associated with scholarly productivity (e.g., publication record, time to degree), emerged as related to interest in faculty careers for the participants in this study. This is not to say that these factors and learning experiences are unimportant or inconsequential. There is considerable evidence that these variables are important for undergraduates considering research careers (McGee and Keller, 2007; Balster et al., 2010; Junge et al., 2010) and for doctoral degree success and completion (Tenenbaum et al., 2001; Anthony and Taylor, 2004; Johnson et al., 2007). However, they were not salient in shaping the career aspirations of the PhD scientists in this study. Instead, alignment of personal values with career opportunities and the structural dynamics of the biomedical academic workforce (i.e., high number of PhDs relative to available academic jobs, pay, availability of grant funding, extremely high workload of faculty) played the most central roles in shaping their career interests.

SCCT provided a helpful lens for understanding much of the way in which this process functions for PhD graduates. For the scientists in this sample, choice actions with respect to training decisions were often not tightly coupled to specific careers or professions. In other words, decisions to pursue a PhD or postdoctoral training were generally not conflated with clear, well-informed career intentions or indicative of the desire to pursue a faculty career. Because these scientists began their PhD training with malleable career interests, their training environments played a strong role in shaping their professional aspirations. However, participants described their graduate and postdoctoral training environments as largely lacking structured career development. Although a number of institutions have in recent years expanded the career development activities offered to trainees, for the vast majority of scientists in this sample, there were neither opportunities to learn about the career options for PhD scientists, nor intentional opportunities to develop relevant competencies beyond their technical skills. Incorporating structured career development activities as a formal part of the graduate curriculum and postdoctoral training would enhance the career outcomes of all scientists—whether or not they pursue academic careers. The lack of structured career development meant that vicarious learning—that is, learning by observation—played a strong role in shaping their outcome expectations and perceptions of faculty positions. These vicarious learning experiences guided outcome expectations about the process of becoming a faculty member and the nature of faculty life.

Importantly, outcome expectations were consistent across social identity and level of interest in faculty careers. Most participants noted that obtaining a faculty position was extremely difficult due to, among other factors, the large supply of PhDs relative to the number of available positions, and that faculty at research institutions face extremely demanding workloads due to, among other factors, decreasing grant success rates (Kaiser, 2012b, 2013). Furthermore, these outcome expectations were not always directly linked to choice goals; those who had high and low interest in faculty careers had similar views of the process of becoming a faculty member and of faculty life. Thus, it seems that outcome expectations, in and of themselves, did not deter or encourage participants to pursue academic careers.

Rather, personal values were the primary lens through which these scientists engaged their outcome expectations as they related to pursuing faculty careers. Previous work has identified barriers, such as research skill and efficacy, inadequate mentoring, and hostile training environments, as having negative implications for career development, particularly for scientists who are women and/or from URM backgrounds (Adams, 1992; Meyers and Turner, 2000; Anthony and Taylor, 2004). Having completed their PhDs, participants in this study had by definition successfully navigated any such barriers. In forming interest in faculty careers, these scientists assessed whether the outcomes they expected in a faculty position were congruent with what matters to them, and the alignment of personal values with opportunities available as a faculty member played a strong role in promoting interest in faculty careers. However, values were not universal; they varied largely by social identity. Thus, our findings suggest personal values are an important additional factor that must be considered in models of career development.

Importantly, beyond personal values, the structural dynamics of the biomedical workforce—long training periods, low postdoctoral pay, tight grant funding, and a tight job market due to a large supply of PhD scientists without a comparable number of opportunities for independent research careers (Stephan, 2012)—were also cited as shifting some scientists’ interest away from faculty careers. The potential implications of structural imbalances in the biomedical workforce have been discussed for at least the past two decades (NRC, 1994, 1998; Teitelbaum, 2008), with concerns that the landscape could cause young scientists to leave research careers because of a lack of attractiveness. Indeed, the scientists in this sample with low interest in academic careers illustrate that this concern may have been well founded, and the phenomenon of talented scientists moving away from biomedical research careers due to systemic issues may already be manifesting itself. The role of structural elements in driving career interests away from faculty careers suggests that the dynamics at universities and in the broader biomedical workforce may themselves be exerting a selective pressure on career choice. These scientists’ narratives suggest that, in the current landscape, developing and maintaining interest in a faculty career for a scientist from any background would require, at a minimum, a willingness to accept a significant decoupling of effort and performance from obtaining the goal due to a supersaturated labor market. Further it would require being an individual who does not foresee the demands of an academic career as an impediment to family/household concerns, who does not see many years of postdoctoral salary as a burden to continued participation, and whose personal values align significantly with the traditional academic reward structure. Additionally,
for many scientists from groups underrepresented in the professoriate, pursuing an academic career requires an ability to experience and persevere through negative experiences or expectations perceived to result from one’s social identity.

Thus, strengthening the professoriate through diversity requires going beyond the focus traditionally placed on increasing skills, numbers of degrees granted, and mentoring opportunities for scientists from underrepresented groups. Individual-based interventions may be necessary, but these data strongly suggest they will in themselves be insufficient. These efforts, along with efforts to improve the training of all early-career scientists, should incorporate a career-pathways perspective, recognizing all five important milestones on the path to independence for a PhD biomedical scientist: 1) the decision to pursue a PhD, 2) graduate training experiences, 3) the decision to pursue postdoctoral training, 4) postdoctoral training experiences, and 5) the career decision after completing postdoctoral training (Figure 2). In particular, transitions into and through postdoctoral training are of importance, as this is a nearly uniform requisite for a faculty position and, at this training step, the participation of women and scientists from URM backgrounds drops significantly from levels seen at the PhD milestone (Martinez et al., 2007; NRC, 2011; National Science Board, 2012).

Moreover, efforts to diversify the nation’s faculties must also be coupled with structural reforms at institutions and throughout the biomedical workforce, addressing the contexts in which scientists operate. A number of reports have presented strategies to increase participation of women in the professoriate, and funding agencies and institutions should use their resources to catalyze the cultural and structural changes at universities that will promote the full participation of female scientists (NAS, 2007; Goulden et al., 2009). The men from URM backgrounds in this study raise the possibility that improved postdoctoral salary and rates of grant funding for biomedical researchers from all backgrounds would go a long way to making faculty careers more attractive to them. Strategies to increase URM representation in the professoriate must consider that women receive the majority of BMS PhDs awarded to scientists from URM backgrounds; therefore, increasing representation of minority faculty must take into account how the intersections of race and gender impact career attainment (Malcom and Malcom, 2011; Ong et al., 2011).

While there are many valuable careers in which PhD scientists can use their skills, improving faculty diversity remains an imperative. If universities and funding agencies want diversity with respect to the scholars filling their faculty ranks, then there must be clear pathways to employment, equitable and inclusive hiring practices, and faculty positions and work environments that are attractive to a more diverse set of scientists. Generating a love of research or confidence in one’s ability to do the work is not enough. Rather, increasing interest in the professoriate among a more diverse range of scholars appears to be a matter of ensuring the professoriate allows greater opportunities to align one’s values with one’s academic work, and structural changes that allow early-career scientists to support their families financially and emotionally on the elongated pathway to independence. Re-thinking graduate education, postdoctoral training, and the nature of faculty positions in these ways may reshape career development, yield a more diverse faculty in years to come, and ultimately increase innovation and the capability for discovery.

ACKNOWLEDGMENTS

The authors thank Yolanda George, Gerard Boulin, and Cathy Ledec (American Association for the Advancement of Science [AAAS] Directorate for Education and Human Resources) and Carr Thompson (Burroughs Wellcome Fund) for resource support of this work, and Donna Vogel and Audrey Trapp (Johns Hopkins University), Justine Johnson (University of Maryland, Baltimore County), and Chinonye Nnakwe (University of Chicago) for additional support. We also thank Sherilllyn Black, Kate Stoll, Heather Metcalf, Richard Linton, and Valerie Wilson for review of the manuscript prior to submission. K.D.G. thanks Muriel Poston, Sylvia James, and Jessie DeAro at the NSF for support of this work. K.A.G. thanks the Pennsylvania State University and University of Maryland for financial support of this research. K.D.G. is a member of the board of directors for the National Postdoctoral Association (NPA). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors alone and do not necessarily reflect the views of the NSF, AAAS, NPA, or any other organization with which the authors are affiliated.
REFERENCES


Martinez ED et al. (2007). Falling off the academic bandwagon. Women are more likely to quit at the postdoc to principal investigator transition. EMBO Rep 8, 977–981.


There are widespread aspirations to focus undergraduate biology education on teaching students to think conceptually like biologists; however, there is a dearth of assessment tools designed to measure progress from novice to expert biological conceptual thinking. We present the development of a novel assessment tool, the Biology Card Sorting Task, designed to probe how individuals organize their conceptual knowledge of biology. While modeled on tasks from cognitive psychology, this task is unique in its design to test two hypothesized conceptual frameworks for the organization of biological knowledge: 1) a surface feature organization focused on organism type and 2) a deep feature organization focused on fundamental biological concepts. In this initial investigation of the Biology Card Sorting Task, each of six analytical measures showed statistically significant differences when used to compare the card sorting results of putative biological experts (biology faculty) and novices (non–biology major undergraduates). Consistently, biology faculty appeared to sort based on hypothesized deep features, while non–biology majors appeared to sort based on either surface features or nonhypothesized organizational frameworks. Results suggest that this novel task is robust in distinguishing populations of biology experts and biology novices and may be an adaptable tool for tracking emerging biology conceptual expertise.

INTRODUCTION

In undergraduate biology education in the United States, there have been national discussions and policy efforts to articulate what an undergraduate biology major should be able to do upon finishing a college degree in the biological sciences (e.g., Association of American Medical Colleges and Howard Hughes Medical Institute [AAMC-HHMI], 2009; Wood, 2009; Labov et al., 2010; Woodin et al., 2010; American Association for the Advancement of Science [AAAS], 2011). While there is no disagreement that students should emerge with more knowledge in biology, goals for students go far beyond an accrual of more information. Specifically, the need to learn to think conceptually like a biologist seems to be a key point of agreement among a variety of stakeholders. The collaboratively published Vision and Change for Undergraduate Biology Education document describes collective aspirations for what many undergraduate biology faculty want students to be able to do upon graduation with a biology degree (AAAS, 2011). The aspirations are lofty, as represented in the following excerpt:

Biology in the 21st century requires that undergraduates learn how to integrate concepts across levels of organization and complexity and to synthesize and analyze information that connects conceptual domains. (p. ix)

In terms of the conceptual domains referred to in this excerpt, Vision and Change goes on to articulate and specify that all undergraduates should develop a basic conceptual understanding of the following core biological concepts:
1) evolution, 2) structure and function, 3) information flow, exchange, and storage, 4) pathways and transformations of energy and matter, and 5) systems (AAAS, 2011).

Taken together, these aspirations represent a desire for biology students to emerge from their undergraduate learning experiences, not simply having accrued a collection of biological information, but also being able to organize and use this information in ways that would resemble “thinking conceptually like a biologist.” So how might we assess whether students are thinking conceptually like a biologist? To what extent could we specifically measure whether students are using the asserted core biological concepts to organize their biological knowledge? And what, if any, evidence is there that diverse populations of biology experts would actually operate using this framework themselves? In the biological sciences, a wonderful array of assessment tools have been developed to assist faculty and administrators in understanding what students are and are not learning about biology as a result of their undergraduate studies (e.g., Odom and Barrow, 1995; Anderson et al., 2002; Baum et al., 2005; Wilson et al., 2007; Bowling et al., 2008; Nehm and Schonfeld, 2008; Smith et al., 2008; Marbach-Ad et al., 2009; Shi et al., 2010; Fisher et al., 2011; Hartley et al., 2011). However, these tools are often targeted at detecting the presence or absence of particular pieces of biological knowledge or misunderstandings about these ideas. To our knowledge, there are currently no assessment tools in biology education that aim to characterize how individuals organize and connect the biological conceptual information they possess.

Cognitive and developmental psychologists, however, have considered the issue of measuring the organization of disciplinary knowledge across many disciplines for decades in studies of what they term expertise. While this term has a variety of meanings in common language, psychologists use “expertise” to represent not only the collection of knowledge that an individual possesses about his or her discipline, but also how that individual organizes and connects that disciplinary knowledge (Bedard and Chi, 1992). In psychological terms, accrual of a large body of information about a discipline is not thought to be sufficient for the development of expertise in that discipline. Rather, that body of information must be organized conceptually in such a way to make retrieval and use of such a large body of information facile in a variety of novel situations (Newell and Simon, 1972; Chi, 2006a).

Cognitive and developmental psychologists have used a variety of research methods to measure and characterize expertise in chess, mathematics, and radiogram reading among radiologists, to name just a few examples (Chi, 2006b). A particularly promising methodological approach across all these studies has been to develop a discipline-based task, often a sorting task, in which the performance of putative experts and novices in organizing examples of disciplinary knowledge can be compared. One such study in physics education offers an approach to gauging the development of expertise in physics (Chi et al., 1981). In this study, participants were asked to sort and categorize physics problems taken from the end-of-chapter sections of a commonly used introductory undergraduate physics textbook. In their study, the researchers asked eight advanced physics doctoral students—classified as “experts”—and eight undergraduates who had completed an introductory course in mechanics—classified as “novices”—to sort 24 physics problems on the basis of similarity of solution. Their results strikingly revealed that experts appeared to group problems on the basis of their underlying conceptual features (e.g., Newton’s laws), whereas novices appeared to group problems on the basis of superficial, contextual features (e.g., blocks on inclined planes). These different sorting results by disciplinary novices and disciplinary experts is hypothesized to reflect differences in how these individuals are mentally organizing the disciplinary knowledge they possess. These organizational frameworks are not thought to be necessarily explicitly or consciously recognized, even by experts themselves (Dreyfus and Dreyfus, 2005). Taken together, these studies suggest that performance on structured tasks such as a card sorting task may have the unique potential to reveal information regarding the structure of an individual’s disciplinary knowledge and, subsequently, his or her level of conceptual expertise in that discipline. Card sorting tasks could, then, be a promising assessment tool in monitoring the nature and development of expertise—referred to above as “thinking conceptually like a biologist”—among individuals at different stages of education or training within a discipline.

While a card sorting task for measuring biological conceptual expertise has not been previously developed, at least one prior investigation suggests that such an approach may reveal differences among biology experts and biology novices (Smith and Good, 1984). In their research aimed at understanding problem-solving approaches used by novices and experts in biology, these researchers interviewed undergraduate students, graduate students, and biology instructors as they solved problems in Mendelian genetics. While the conclusions of their paper focused primarily on the great variety of problem-solving strategies they documented, Smith and Good made another key observation “that unsuccessful subjects tend to categorize problems according to their superficial characteristics (e.g., a flower problem) instead of deeper features (e.g., a monohybrid problem).” These results suggest that an adaptation of the physics problem-sorting study could produce an assessment task on which novices and experts in biology may perform differently. The development of such a biology card sorting task could potentially fill a gap in the assessment tool portfolio that is currently available to biology education practitioners and researchers, providing a novel assessment tool to gauge conceptual expertise in biology.

The purpose of this study was to adapt measurement approaches from cognitive and developmental psychology (Chi et al., 1981) to develop an assessment tool in biology education that could be used to measure conceptual expertise in biology. In this paper, we present that novel assessment tool, the Biology Card Sorting Task. This task has been designed to probe how an individual organizes his or her conceptual knowledge of biology. The primary aim of this initial study of the Biology Card Sorting Task was to test the hypothesis that putative biological experts (biology faculty) and putative biological novices (non—biology majors) would perform differently on this sorting task, using a variety of quantitative metrics. We describe the unique structure of this card sorting task, novel quantitative methods developed to analyze the resulting data, and our initial findings in using this Biology Card Sorting Task as a biology education research and assessment tool.
METHODS

Building upon the work of Chi et al. (1981) in physics education research, a novel card sorting task was designed to distinguish different levels of biological expertise and to compare the nature of biological expertise of different participant populations. By exploring the ways in which biology faculty and undergraduate non-biology majors arranged biology problems into groups, justified their arrangement, and named the groups, we probed how they organized their knowledge of fundamental biological concepts. In this section, we describe the development of the card sorting task, the implementation of the task, and the multiple new analytical approaches developed to quantify card sorting differences within and across participant populations with regard to: constructed card groupings, constructed card group names, and responses to reflective prompts. Finally, we describe recruitment of the participant population for this initial investigation of the Biology Card Sorting Task.

Task Development

Prior to the research presented here, we conducted a small-scale pilot study. Twenty-six biology problems were taken unsystematically from commonly used introductory biology curricula and printed on cards. Subjects \( n = 122 \) undergraduate students in an upper-division biology course were asked to sort the problems into groups representing fundamental biological principles. The results from this preliminary study were unwieldy and provided a key insight: a robust analysis and interpretation of card sorting data as a measure of biological expertise would require proposing and testing specific organizational frameworks that individuals may be using to organize their biological knowledge. This, in turn, would require the development of a hypothesis-driven card stimulus set that would be the basis of the card sorting task.

We hypothesized that biological novices would be most likely to sort biology problems based on the surface feature of organism type and that biological experts would be most likely to sort based on deep features of the problems, namely, core biological concepts. Subsequently, 16 biology problems were specifically selected to provide a card stimulus set based on these two hypothesized organizational frameworks: one that was based on four surface features (organism types; see row titles in Figure 1) and another that was based on four deep features (core biological concepts; see column titles in Figure 1). Given this card stimulus set, our working hypothesis was that novices would construct four card groupings representing four surface features: 1) the card grouping K, D, J, and I represented the surface feature “Plant”; 2) the card grouping H, F, B, and M represented the surface feature “Insect”; 3) the card grouping N, L, O, and P represented the surface feature “Human”; and 4) the card grouping C, A, E, and G represented the surface feature “Microorganism” (Figure 1). We hypothesized that experts presented with this same card stimulus set would construct four orthogonal card groupings representing four deep feature categories: 1) the card grouping K, H, N, and C represented the deep feature “Evolution by Natural Selection in Living Systems”; 2) the card grouping D, F, L, and A represented the deep feature “Pathways and Transformations of Energy and Matter in Living Systems”; 3) the card grouping J, B, O, and E represented the deep feature “Storage and Passage of Information about How to Build Living Systems”; and 4) the card grouping I, M, P, and G represented the deep feature “Relationships between Structure and Function in Living Systems” (Figure 1). These four deep feature categories were aligned with the core concepts of biological literacy highlighted in Vision and Change and the recently revised AP Biology curriculum framework (AAAS, 2011; College Board, 2013). The fifth biological concept described in Vision and Change, systems, was not represented as a separate category, but instead was integrated into the titles of the other four core concepts.

The 16 biology problems used were chosen from four widely used curricular sources (Udovic et al., 1996; Hickman et al., 2007; Campbell et al., 2008; Raven et al., 2011) and were edited to enhance readability and to eliminate jargon, graphics, and direct word cues related to core biological concepts. The actual problems used in the study are not included for publication to limit access to students and maintain the integrity of the task, but they may be obtained by a request to the senior author (K.D.T.).

Task Conditions

We used this hypothesis-driven card set to probe how non-biology undergraduate majors and biology faculty organize biological concepts. As this was intended to be a conceptual rather than a knowledge-based task, participants were asked to read the problems and explicitly instructed that they were not to solve them. Participants were also told that the task was not intended as a test and that there were no right or wrong ways to organize the cards. Participants were given as much time as they deemed necessary to complete each task. Non-biology majors completed the tasks during the first laboratory meeting of their course. Biology faculty completed the tasks in a one-on-one format led by a member of the research team (J.I.S.) who was unfamiliar to most of the faculty and not a member of their department. Each subject was asked to complete two card sorting tasks: first in the unframed condition and then in the framed condition (Figure 2). Each of these task conditions is described below.

Table 1. Hypothesized biology card stimulus set. Columns and rows represent each of the four hypothesized deep features of biology, and letters represent each of the four hypothesized surface features of biology. Each letter represents one of sixteen biology problems that was printed on a card with that letter at the top. Each problem was crafted to contain a single hypothesized deep feature and a single hypothesized surface feature.
Unframed Sorting Task Condition. In the unframed condition (Figure 2A), participants were asked to consider what they knew about biology and to sort the problems into groups representing common underlying biological principles. Each subject was allowed to decide for himself or herself what that meant. Subjects were encouraged to generate more than one group and fewer than 16 with the proviso that a single problem could not be a member of more than one group. Subjects were asked then to name each group of cards in a way that described what the commonality of the group was for them. Participants recorded groupings, the number of groups, names of the groups, and start and stop times on a form provided by the researchers. After completing the unframed card sorting, subjects were then asked to respond to two reflective prompts that probed the reasoning behind their card groupings and group names: 1) “Describe why you grouped certain problems together and give an example of your reasoning.” 2) “How did you decide on the names of your groups?”

Framed Sorting Task Condition. After completion of the unframed sorting task, the framed card sorting task (Figure 2B) was used to determine the extent to which participants could sort the problems into the four hypothesized deep feature categories when explicitly cued by these category names. In the framed sorting task condition, participants were asked to sort the 16 problems again, but this time into four groups that had been preassigned the following names by the researchers: 1) “Evolution by Natural Selection in Living Systems,” 2) “Pathways and Transformations of Energy and Matter in Living Systems,” 3) “Storage and Passage of Information about How to Build Living Systems,” and 4) “Relationships between Structure and Function in Living Systems.” Participants were asked to record the cards they placed into each group under the given headings, as well as start and stop times, on a form provided by the researchers. When finished sorting, subjects were then asked to respond to two reflective prompts: 1) “Which if any of the problems was difficult to assign to one of the 4 categories and why? Please list all that apply.” 2) “Now that you have completed 2 card sorting activities, which group names do you prefer: the group names that you created or the group names given to you by the researchers or neither? Please explain your answer.”

To assist the researchers in making comparisons among different groups, we asked participants to respond to a variety of questions regarding themselves and their educational background after completing both of the sorting tasks. Demographic information regarding gender, ethnicity, and major field of study of the participant populations will be reported here.

Analysis and Comparison of Constructed Card Groupings
Participants may have organized their cards based upon hypothesized surface features (organism type), hypothesized deep features (core biological concepts), or some other unexpected sorting scheme. To quantify how similar the card groupings generated by participants were to our hypothesized groupings (Figure 1), we developed two analytical metrics to describe each sort: percent card pairings and edit distance.

Percent Card Pairings. Percent card pairings measured the degree to which the cards grouped by a participant generated pairings predicted as surface feature pairings, deep feature pairings, or unexpected pairings. For example, in the card group [CDK], one card pair—CK—belongs to the hypothesized deep feature group, “Evolution by Natural Selection in Living Systems” (Figure 1). Another pair—DK—belongs to the surface feature group “Plant” (Figure 1). The final card pair—CD—represents an unexpected pairing; it belongs to neither the hypothesized surface nor the deep feature groupings (Figure 1). Our card stimulus set contains 24 hypothesized deep feature pairings, 24 hypothesized surface feature pairings, and 88 unexpected pairings. If a participant generated a group with a single card, then it was treated as an unexpected pairing. Percentages of deep feature, surface feature, and unexpected card pairings were calculated for each participant by identifying all the card pairs within each card group generated by the participant. These calculations were performed using a card entry Python (Python Software Foundation, 2011) script written by the researchers that generated a spreadsheet of all the card pairings contained in each sort. The number of deep feature, surface feature, and unexpected pairings for each individual for each of his or her sorts was counted using Microsoft Excel (Redmond, WA). Percentages of deep feature pairings, surface feature pairings, and unexpected pairings were averaged across individuals in each participant population for both the unframed and framed conditions, and were then compared.

Edit Distance. A second analytical approach was used to quantify and compare sorting results. Edit distance (Deibel et al., 2005) was measured as the minimum number of card moves needed to turn an individual’s card sort either into an exact hypothesized surface feature sort or an exact hypothesized deep feature sort. For example, the set of card groups [{ADL}, {BEFJ}, {O}, {CGHK}, and {IMP}] could be converted into the hypothesized deep feature sort [{ADFL}, {BEJO}, {CHKN}, and {GIMP}] by moving three cards: F, G, and O. Using this approach, an edit distance from the exact hypothesized surface feature sort (ED-Surface) and an edit distance from the exact hypothesized deep feature sort (ED-Deep) could be calculated for each individual card sort. As such, an exact hypothesized deep feature sort would have an ED-Deep of 0 necessary card moves and ED-Surface of 12 necessary card moves. Similarly, an exact hypothesized surface feature sort would have an ED-Deep of 12 necessary
card moves and ED-Surface of 0 necessary card moves. Units of ED are “necessary card moves” and were calculated using the Hungarian method (Kuhn, 1955) and Clapper’s (2008) munkres implementation written in Python script by the researchers. From these calculated edit distances, an average ED-Surface and ED-Deep were calculated for and compared between the non–biology major and the biology faculty populations, as well as between the unframed and framed task conditions for each population.

**Analysis and Comparison of Constructed Card Group Names**

A scoring rubric was developed to determine the extent to which group names given by the participants in the unframed condition matched hypothesized surface features and hypothesized deep features. The scoring rubric was revised using subsets of the data until at least 90% interrater reliability was achieved. Group names given by participants that did not match hypothesized features were not included in this analysis. The percentage of participants in each population that gave group names similar to the hypothesized surface features and deep features were calculated and then compared.

**Analysis and Comparison of Responses to Reflective Prompts**

A scoring rubric was also developed to assess the prevalence of sorting strategies based upon surface or deep features in a combined analysis of all the responses given by participants to the four reflection questions. In addition, we examined the prevalence of participants preferring the names that they gave to the card groupings (in the unframed condition) to those preferring the names provided by the researchers (in the framed condition).

**Comparative Statistical Analyses**

Two-tailed Student’s t-tests were used to compare the average percent card pairing and average edit distance measures between participant populations within a task condition. Comparisons of the results for a single participant population between the two task conditions—unframed and framed—were similarly analyzed. Additionally, we used a 2 (group: biology faculty, non–biology majors) by 2 (task: unframed, framed) repeated-measures analysis of variance (ANOVA) to examine the significance level of all of these comparisons. Results from t-test analyses were confirmed by repeated-measures ANOVAs, so only the former are presented here. Pearson’s chi-square tests were used to compare the prevalence of group names and specific card sorting strategies used by different participant populations. To normalize for differences in the size of particular participant populations, all variances are presented as an SE of the mean. All statistical comparisons were generated using JMP, version 9 (2010), or IBM SPSS, version 20.0 (2011).

**Recruitment and Participant Population**

Participants in this research were recruited from the students and faculty of a large urban university with more than 25,000 undergraduates (1800 biology majors and 5000 students enrolled in biology courses) and ~40 faculty members in biology who are active in research, as well as in teaching, and represent a wide breadth of subdisciplines spanning from the molecular to the ecological scale. Biology experts were recruited from among the tenured and tenure-track faculty. Biology novices were recruited from a general education course in biology populated primarily by non–biology major students during the first week of the semester. We hypothesized that non–biology majors enrolled in a biology course would have the greatest interest in and understanding of biology among the population of non–biology majors on campus, and thus we thought this population would be most representative of university-level biology novices. Each student in the course completed the tasks associated with this study as part of his or her course curriculum, but only those identified as non–biology majors were included in the study. Likewise, only those participants who completed tasks as directed by the researchers were included in the study. Each subject was allowed to decline participation without negative consequence. The committee for the protection of human subjects approved this research (protocol #X10-036).

**RESULTS**

This Biology Card Sorting Task yields multiple sources of data for analysis. To evaluate this task as a novel assessment tool, it is important to consider multiple measures of participants’ performance. Below, a description of the participant populations is followed by example raw card sort data. Then, six analyses are presented that provide insights not only into how the two subject populations group the cards in the stimulus set, but also how these two populations named their constructed card groupings and rationalized their approach to the task. The six analyses presented are: 1) analysis of the number of card groups generated in the unframed task condition; 2) analysis of time to sort in both the unframed and framed task conditions; 3) calculation of prevalence of deep feature, surface feature, and unexpected card pairings in both task conditions; 4) calculation of edit distance from hypothesized deep feature and surface feature sorts in both task conditions; 5) analysis of prevalence of hypothesized deep feature and surface feature group names in the unframed task condition; and 6) analysis of card sorting rationales based on responses to the reflective questions. The figures, tables, and results are organized to show comparisons between non–biology majors and biology faculty and comparisons between the unframed and the framed card sort task conditions for each of these participant populations.

**Description of Participant Populations**

The participant populations for this biology card sort study are described in Table 1. From an invited pool of 35 tenured/tenure-track biology faculty, 23 participated for a 69% participation rate. From an invited pool of 131 undergraduate non–biology majors, all participated in doing the task as part of their class activity and the results for 101 are presented here for an actual 89% participation rate. Of the 30 undergraduates whose data are not included here, two were excluded because they were actually biology majors, 14 did not consent to have their data included in the study, and
Table 1. Participant population

<table>
<thead>
<tr>
<th>Participant type</th>
<th>Number invited</th>
<th>Participation rate</th>
<th>Sample size</th>
<th>Female participants</th>
<th>Participants of color</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tenure-track biology faculty</td>
<td>35</td>
<td>69%</td>
<td>23</td>
<td>26%</td>
<td>48%</td>
</tr>
<tr>
<td>Undergraduate non–biology major</td>
<td>131</td>
<td>89%</td>
<td>101</td>
<td>55%*</td>
<td>55%</td>
</tr>
</tbody>
</table>

*p = 0.014 (Pearson chi-square).

14 had sorting anomalies (e.g., using a card twice or failure to use a card). There was a significantly greater proportion of females in the non–biology major population (55%) compared with the biology faculty population (26%; $\chi^2 = 6.032, p = 0.014$). There was no significant difference in the proportion of participants of color between non–biology majors (55%) and biology faculty (48%). All analyses and comparisons described below are based on these non–biology majors ($n = 101$) and biology faculty ($n = 23$).

Example Card Sorts from a Non–Biology Major and a Biology Faculty for the Unframed and the Framed Task Conditions

Figure 3 shows example biology card sort task results from a single biology faculty member (Figure 3, A and B) and a single non–biology major student (Figure 3, C and D). These examples are shown to highlight two things. First, it is important to note that the biology card sorting task yields multiple sources of data for analysis, including the number of card groups generated, the membership and resulting card associations that result from these constructed groups, and the chosen name that an individual assigns to each group in the unframed task condition. Data from the reflective questions were also analyzed (raw data not shown). Second, these examples are shown to demonstrate that an exact hypothesized deep feature card sort (see columns in Figure 1) and an exact hypothesized surface feature card sort (see rows in Figure 1) were observed in the unframed task condition (see Figure 3, A and C, respectively). Of note, no biology faculty ever produced an exact hypothesized surface feature sort, nor did any non–biology major ever produce an exact hypothesized deep feature sort in the unframed task condition.

Analysis of Card Groupings and Resulting Surface Feature, Deep Feature, and Unexpected Card Pairings Constructed by Non–Biology Majors and Biology Faculty

As described in Methods, comparisons between constructed card groupings were accomplished by identifying all the card pairs that existed within a card group for each of the groups...
generated by an individual participant. Then, a percentage of hypothesized deep feature card pairs, hypothesized surface feature card pairs, and unexpected card pairs could be calculated for each individual’s card sort. This approach was used to calculate average percentages of different types of card pairs for non–biology major and biology faculty populations, as well as to compare these proportions of different types of card pairs in the unframed and framed task conditions.

**Percent Card Pairings in the Unframed Card Sort.** In the unframed card sort condition (Figure 4A and Table 2), Biology faculty \((n = 23)\) generated an average of 8.6 ± 2.2% surface feature card pairings, 71.7 ± 3.9% deep feature card pairings, and 19.8 ± 2.6% unexpected card pairings. In the unframed card sort, non–biology majors \((n = 101)\) generated an average of 40.8 ± 2.9% surface feature card pairings, 29.2 ± 2.2% deep feature card pairings, and 30.0 ± 1.6% unexpected card pairings. Statistical comparison of these means showed that non–biology majors generated a significantly smaller average percentage of deep feature card pairings in the unframed task condition than biology faculty \((p < 0.0001)\). In addition, non–biology major students generated a significantly greater average percentage of surface feature card pairings \((p = 0.0001)\) and unexpected card pairings \((p = 0.0018)\) in the unframed condition compared with biology faculty.

**Percent Card Pairings in the Framed Card Sort.** In the framed card sort condition (Figure 4B and Table 2), non–biology majors \((n = 101)\) generated an average of 16.2 ± 0.8% surface feature card pairings, 39.6 ± 2.0% deep feature card pairings, and 44.2 ± 1.5% unexpected card pairings. In the framed card sort, biology faculty \((n = 23)\) generated an average of 4.3 ± 0.8% surface feature card pairings, 83.1 ± 3.3% deep feature card pairings, and 12.6 ± 2.5% unexpected card pairings. Statistical comparison of these means showed that non–biology majors continued to generate a significantly smaller proportion of deep feature card pairings in the framed task condition than biology faculty \((p < 0.0001)\). In addition, non–biology major students continued to generate a significantly higher proportion of surface feature card pairings \((p < 0.0001)\) and unexpected card pairings \((p < 0.0001)\) in the framed task condition compared with biology faculty.

**Comparison of Percent Card Pairings between the Unframed and Framed Card Sorts.** Comparison of data from the unframed and framed conditions revealed significant shifts between the two task conditions within each participant population. In the framed task condition, non–biology majors constructed a significantly smaller proportion of surface feature pairs \((16.2 ± 0.8\% \text{ vs. } 40.8 ± 2.9\%;\ p < 0.0001)\) and a significantly greater proportion of deep feature \((39.6 ± 2.0\% \text{ vs. } 29.2 ± 2.2\%;\ p = 0.0006)\) and unexpected pairs \((44.2 ± 1.5\%)\) vs. 30.0 ± 1.6%;\(p < 0.0001\)) compared with their initial results in the unframed task condition. Similarly, in the framed task, biology faculty constructed a significantly greater proportion of deep feature pairs \((83.1 ± 3.3\% \text{ vs. } 71.7 ± 3.9\%;\ p < 0.0291)\) and a somewhat smaller proportion of surface feature pairs

![Table 2](image)

<table>
<thead>
<tr>
<th>Participant type</th>
<th>Unframed task condition</th>
<th>Framed task condition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Surface</td>
<td>Deep</td>
</tr>
<tr>
<td>Tenure-track biology faculty</td>
<td>23</td>
<td>8.6%</td>
</tr>
<tr>
<td>Undergraduate non–biology major</td>
<td>101</td>
<td>40.8%* (2.9)</td>
</tr>
</tbody>
</table>

*p < 0.001 for comparisons between participant types on each measure shown above. SEM is in parentheses.
(4.3 ± 0.8% vs. 8.6 ± 2.2%; p < 0.0849) and unexpected pairs (12.6 ± 2.5% vs. 19.8 ± 2.6%; p < 0.0547) compared with their initial results in the unframed task condition. In summary, both participant populations shifted toward a significantly greater proportion of deep feature card pairings, but only non–biology majors showed a significantly smaller proportion of surface feature card pairs and a significantly greater proportion of unexpected pairs in the framed task condition compared with the unframed task condition.

**Analysis of Edit Distances from the ED-Surface and the ED-Deep Sorts**

As described in *Methods*, comparisons between constructed card groupings by non–biology majors and biology faculty was also accomplished through a second analysis, in which an edit distance from the ED-Surface sort and an edit distance from the ED-Deep sort was calculated for each individual card sort. Edit distance is defined as the minimum number of card moves that would need to be made to turn an individual’s card sort either into an exact hypothesized surface feature sort or an exact hypothesized deep feature. As such, an exact hypothesized deep feature sort would have an ED-Deep of 0 necessary card moves and ED-Surface of 12 necessary card moves. Similarly, an exact hypothesized surface feature sort would have an ED-Deep of 0 necessary card moves and ED-Surface of 12 necessary card moves. From these calculated edit distances, an average ED-Surface and ED-Deep could be calculated and compared for the non–biology major population and the biology faculty population, as well as between the unframed and framed task conditions for each population. Units of ED are “necessary card moves.”

**Edit Distance in the Unframed Card Sort.** In the unframed card sort condition (Figure 5A and Table 3), biology faculty (*n = 23*) constructed card sorts with an average ED-Surface of 10.9 ± 0.3 and an average ED-Deep of 4.5 ± 0.5. Non–biology majors (*n = 101*) constructed card sorts with an average ED-Surface of 7.0 ± 0.3 and an average ED-Deep of 8.2 ± 0.3. Statistical comparison of these means showed that non–biology majors and biology faculty were significantly different from one another (*p < 0.0001)*.

**Edit Distance in the Framed Card Sort.** In the framed card sort condition (Figure 5B and Table 3), non–biology majors (*n = 101*) constructed card sorts with an average ED-Surface of 9.1 ± 0.1 and an average ED-Deep of 5.9 ± 0.3. Biology faculty (*n = 23*) constructed card sorts with an average ED-Surface of 11.0 ± 0.2 and an average ED-Deep of 1.2 ± 0.3. Statistical comparison of these means showed that non–biology majors and biology faculty were significantly different from one another (*p < 0.0001)*.

**Comparison of Edit Distance between the Unframed and Framed Card Sorts.** Comparison of ED-Surface and ED-Deep analyses for the unframed and framed task conditions

<table>
<thead>
<tr>
<th>Participant type</th>
<th>Unframed task condition</th>
<th>Framed task condition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>ED from surface sort</td>
</tr>
<tr>
<td>Tenure-track biology faculty</td>
<td>23</td>
<td>10.9 (±0.3)</td>
</tr>
<tr>
<td>Undergraduate non–biology major</td>
<td>101</td>
<td>7.0* (±0.3)</td>
</tr>
</tbody>
</table>

*p < 0.0001 for comparisons between participant types on each measure shown above. SEM is in parentheses.

*Note that lower ED numbers indicate sorts more similar to the hypothesized sort.*
revealed significant shifts between the two task conditions for each participant population. In the framed task, non–biology majors constructed card sorts that had ED-Surface ($9.1 \pm 0.1$) that were statistically further away from an exact hypothesized surface feature sort compared with their constructed sorts in the unframed task condition ($7.0 \pm 0.3; p < 0.0001$). In addition, non–biology majors constructed card sorts in the framed task condition that had ED-Deep ($5.9 \pm 0.3$) that were significantly closer to an exact hypothesized deep feature sort compared with their initial results in the unframed task condition ($8.2 \pm 0.3; p < 0.0001$). In the framed task, biology faculty constructed card sorts that had ED-Surface ($11.0 \pm 0.2$) that were statistically indistinguishable from their constructed sorts in the unframed task condition ($10.9 \pm 0.3; p = 0.6827$). However, biology faculty constructed card sorts in the framed task condition that had ED-Deep ($1.2 \pm 0.3$) that were significantly closer to an exact hypothesized deep feature sort compared with their initial results in the unframed task condition ($4.5 \pm 0.5; p < 0.0001$). In summary, both participant populations shifted toward constructing card groupings that were more similar to the exact hypothesized deep feature sort in the framed task condition compared with the initial conditions, with non–biology majors also significantly shifting away from constructing card groupings closer to the exact hypothesized surface feature sort.

**Analysis and Comparison of Constructed Card Group Names**

While the analyses presented above provide insights into how the participant populations grouped the cards in the stimulus set, the analyses below describe how these participant populations chose to name their constructed card groupings in the unframed task condition. As described in Methods, comparisons between the names given to constructed card groupings by non–biology majors and biology faculty were accomplished through blind coding of the card group names.

<table>
<thead>
<tr>
<th>Hypothesized Deep Feature Group Names</th>
<th>Titles generated by subjects in the unframed sort that were accepted as equivalent to those given by researchers in the framed sort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evolution by natural selection in living systems</td>
<td>Any title using the term <em>evolution</em></td>
</tr>
<tr>
<td>Pathways and transformations of energy and matter in living systems</td>
<td>Any title using the term <em>energy</em></td>
</tr>
<tr>
<td>Storage and passage of information about how to build living systems</td>
<td>Any title using the term <em>DNA</em> or <em>traits</em></td>
</tr>
<tr>
<td>Relationships between structure and function in living systems</td>
<td>Any title using the term <em>form</em> or <em>function</em></td>
</tr>
</tbody>
</table>

For the presence of hypothesized deep features or hypothesized surface features. Two observers analyzed all card group names. Groups names that included the terms *human, plant, insect,* or *microorganism* were coded as representing each of those hypothesized surface features, respectively. For deep features, it was necessary to develop a rubric more specifically defining the terminology found in group names that would be coded as representing each hypothesized deep feature (see Table 4). These rubrics were developed and refined based on responses seen in the data set, examples of which are listed in the second column. Interrater reliability for each analysis presented was greater than 95% agreement between observers.

**Hypothesized Deep Feature Group Names** Analysis of the prevalence of card group names related to the four hypothesized deep features (see columns in Figure 1) is shown in Figure 6. For all four hypothesized deep features, a significantly larger proportion of biology faculty ($n = 23$) used each hypothesized deep feature in naming one or more of their card groups as compared with non–biology majors ($n = 101$). More specifically, the deep feature “Evolution by Natural Selection in Living Systems” appeared in the group names of a significantly larger proportion of biology faculty (87.0%) as compared with non–biology majors (19.8%; $\chi^2 = 38.7, df = 1, p < 0.0001$; Figure 6A). The deep feature “Pathways and Transformations of Energy and Matter” also appeared in the group names of a significantly higher proportion of biology faculty (82.6%) as compared with non–biology majors (8.9%; $\chi^2 = 58.2, df = 1, p < 0.0001$; Figure 6B). The deep feature “Storage and Passage of Information about How to Build Living Systems” appeared in the group names of a significantly larger proportion of biology faculty (95.7%) as compared with non–biology majors (39.6%; $\chi^2 = 23.5, df = 1, p < 0.0001$; Figure 6C). Finally, the deep feature “Relationships between Structure and Function in Living Systems” appeared in the group names of 39.1% of biology faculty as compared with non–biology majors (19.8%; $\chi^2 = 23.5, df = 1, p < 0.0001$; Figure 6C).
with none of the non–biology majors (0%; $\chi^2 = 42.6$, df = 1, $p < 0.0001$; Figure 6D).

**Hypothesized Surface Feature Group Names.** Analysis of the prevalence of card group names related to the four hypothesized surface features (see rows in Figure 1) is shown in Figure 7. For all four surface features, a significantly larger proportion of non–biology majors ($n = 101$) used each hypothesized surface feature in naming one or more of their card groups as compared with biology faculty ($n = 23$). More specifically, the surface feature “Human” appeared in the group names of a significantly larger proportion of non–biology majors (47.5%) as compared with biology faculty (8.7%; $\chi^2 = 11.7$, df = 1, $p = 0.0006$; Figure 7A). The surface feature “Insect” appeared in the group names of 43.6% of non–biology majors as compared with none of the biology faculty (0%; $\chi^2 = 15.5$, df = 1, $p < 0.0001$; Figure 7B). The surface feature “Plant” appeared in the group names of a significantly larger proportion of non–biology majors (51.5%) as compared with biology faculty (17%; $\chi^2 = 8.8$, df = 1, $p = 0.003$; Figure 7C). Finally, the surface feature “Microorganism” appeared in the group names of 36.6% of non–biology majors as compared with none of the biology faculty (0%; $\chi^2 = 12.0$, df = 1, $p = 0.0005$; Figure 7D).

**Analysis and Comparison of Card Sorting Strategy Explanations from Responses to Posttask Reflection Questions**

In addition to analyzing how participants grouped the cards and named these groups, we also analyzed participants’ reported card sorting strategy explanations, which appeared in their responses to posttask reflection questions. As described in Methods, analysis of participants’ card sorting strategies was accomplished through blind coding of participants’ reflection narratives. In particular, narratives were coded for rationales that included explicit reference to using either hypothesized surface features or hypothesized deep features in sorting. Two observers analyzed all narrative responses to the reflection questions, and the interrater reliability for each analysis presented was greater than 94% agreement between observers.

Analysis of the participants’ card sorting strategy explanations is shown in Table 5. Sample quotes from both biology faculty and non–biology majors who evidenced each rationale are shown. In their reflection narratives, 100% of biology faculty made reference to one or more of the four hypothesized deep features as part of their sorting strategy as compared with only 22% of non–biology majors ($\chi^2 = 47.9$, df = 1, $p < 0.0001$). In contrast, only two faculty (8.7%) made reference to using hypothesized surface features in their sorting, as compared with 37.6% of non–biology majors ($\chi^2 = 7.2$, df = 1, $p = 0.0074$).

**Analysis of Number of Card Groups Generated in Unframed Task Condition**

The average number of card groups generated in the unframed task by non–biology majors ($5.3 \pm 0.2$, $n = 101$) was significantly fewer than the number generated by biology faculty ($6.5 \pm 0.3$, $n = 23$; $p = 0.0011$).
The Biology Card Sorting Task

Figure 7. Prevalence of surface features in card group names in the unframed task condition. Proportions of participants who included (black bars) or did not include (white bars) each surface feature in one or more of their card group names in the unframed task condition are shown for non–biology majors (left) and biology faculty (right) for each of the hypothesized surface features: (A) “Human,” (B) “Insect,” (C) “Plant,” and (D) “Microorganism.” A significantly larger proportion of non–biology majors included each surface feature in their card group names as compared with biology faculty ($p < 0.003$ or lower for each surface feature). See Results for analysis rubric and statistical details.

Average Completion Times for the Unframed and Framed Task Conditions

The average time taken to complete the unframed task condition by non–biology majors (12.0 ± 0.5 min, $n = 96$) was significantly shorter ($p = 0.0547$) than for biology faculty (15.1 ± 1.5 min, $n = 23$). However, the average time taken to complete the framed task condition by non–biology majors (6.8 ± 0.3 min, $n = 93$) was significantly longer ($p = 0.0001$) than for biology faculty (4.7 ± 0.4 min, $n = 23$). Comparison of task completion time data within each participant population for the unframed and framed conditions revealed significantly faster task completion times for both non–biology majors (12.0 ± 0.5 vs. 6.8 ± 0.3 min; $p < 0.0001$) and biology faculty (15.1 ± 1.5 vs. 4.7 ± 0.4 min; $p < 0.0001$) in the framed task condition.

Table 5. Rubric for and analysis of card sorting strategy explanations

<table>
<thead>
<tr>
<th>Participant type</th>
<th>Surface feature rationale</th>
<th>Sample quote</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tenure-track biology faculty ($n = 23$)</td>
<td>8.7%</td>
<td>“Others [cards] are united by the kind of organism (DIPMG).”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Plant, well plant are plants and they are just odd.”</td>
</tr>
<tr>
<td>Undergraduate non–biology major ($n = 101$)</td>
<td>37.6%</td>
<td>“I grouped certain cards together because I looked for key words in the problems such as: insects, humans, cells, and plants.”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“With plants, any reference to photosynthesis, reproduction, environment, would categorize them into such a group.”</td>
</tr>
</tbody>
</table>

Deep feature rationale

<table>
<thead>
<tr>
<th>Participant type</th>
<th>Surface feature rationale</th>
<th>Sample quote</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tenure-track biology faculty ($n = 23$)</td>
<td>100%</td>
<td>“1st group dealt w/ manner in which all organisms (microbe, plant or animal) process energy within their cells, energy metabolism was a unifying theme.”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Evolution and natural selection make sense to me but in another level I also liked structure and function. As for L, on the surface it seems not to fit but when you attempt to answer the question, energy and matter seemed the best category.”</td>
</tr>
<tr>
<td>Undergraduate non–biology major ($n = 101$)</td>
<td>22%</td>
<td>“Bacteria &amp; Pesticides evolving to battle their ‘cures’ was categorized as evolution.”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Energy seemed to be an underlying theme (atp) for a few.”</td>
</tr>
</tbody>
</table>
The Biology Card Sorting Task Distinguishes Putative Experts and Novices—Differences in How Non–Biology Majors and Biology Faculty Perform on the Task

Biology faculty and non–biology majors differed significantly on each of the six analytical measures for the Biology Card Sorting Task used to compare the card sorts produced by these two groups. These data confirm that this novel assessment task appears to be robust in distinguishing populations of putative biology experts and putative biology novices on a variety of measures. In addition, this initial investigation also revealed multiple interesting findings about the differences in card sorting results for these two populations. Similar to previous studies, biology faculty took more time to complete their sorts than did non–biology majors (Chi et al., 1981). In addition, biology faculty constructed on average one more card group than non–biology majors. More specific results for each population are considered below.

Biology Faculty Appear to Sort Based on Deep Features. Strikingly, biology faculty grouped cards together in a manner we hypothesized experts would, grouping cards primarily based on deep features (fundamental biological principles). This conclusion is supported by the high proportion of deep feature card pairings produced in the unframed task condition by biology faculty (Figure 4A and Table 2). In addition, the observation that biology faculty had average card sort edit distances closer to the hypothesized deep feature sort than the hypothesized surface feature sort is additional supporting evidence (Figure 5A and Table 3). That biology faculty members are sorting based on deep features is also supported by their sorting strategies, with 100% of faculty naming one or more of the deep features as a part of their sorting strategy (Table 5), and by choices in naming the card groups that they create, explicitly choosing group names that aligned well with the four hypothesized deep feature categories (see Results and Figure 6). All these data support the hypothesis that biology faculty members appear to be primarily using deep features—namely core biological concepts—to sort the biology problems on the cards.

Non–Biology Majors Do Not Necessarily Sort Based on Surface Features. As hypothesized, non–biology majors did not appear to group cards based primarily on deep features; however, they also did not appear to group cards based on surface features, as we had hypothesized they would. This was evidenced by the presence of relatively comparable proportions of surface feature, deep feature, and unexpected card pairings (Figure 4A and Table 2), as well as by the fact that non–biology majors produced card sorts in the unframed condition with edit distances that were almost equidistant from both the hypothesized novice and expert card sorts (Figure 5A and Table 3). In examining how non–biology majors named their card groups, 40% of non–biology majors used language aligned with the four surface features (see Results and Figure 7) to describe their card groupings, whereas a range of 0–40% used the four deep feature group names (see Results and Figure 6). These data suggest that the non–biology major population may be sorting using a variety of different organizational frameworks, not just an organismal framework. Multiple hypotheses could explain the variation seen among non–biology majors in this task. First, non–biology majors are a less homogeneous population by nature than faculty. Some non–biology majors may have intermediate biological knowledge and may be performing this assessment task like partial experts, whereas others may have had minimal exposure to biology and may even struggle with the language of the cards. Even though we attempted to minimize biological jargon on the card stimulus set, there were comments by a subset of non–biology majors that some language on a few of the cards was challenging (e.g., bilateral symmetry; stratified epithelium [unpublished data]). In future versions of the card stimulus set, we would further translate or remove that biological language identified by participants in this study as challenging. Finally, published card sorting results among novices across a variety of fields of expertise tend to exhibit wider variation than results found among experts (Chi et al., 1981; Chi, 2006a; Mason and Singh 2011). A future study to conduct think-aloud interviews with non–biology majors may provide additional insights into the variety of organizational frameworks this particular population is using to perform the assessment task.

Biological Framing Sharpens Biology Faculty Sorts toward Deep Features and Shifts Non–Biology Major Sorts Away from Surface Features. When we compare the card sorting results between the unframed and framed task conditions, biology faculty appear to sort very similarly, with some sharpening of their use of a hypothesized deep feature framework. Biology faculty produced more deep feature card pairings and fewer surface feature and unexpected card pairings in the framed task condition (Figure 4 and Table 2); however, only the increase in deep feature pairs was statistically significant. Similarly, the biology faculty edit distance from the hypothesized expert sort was reduced to only ~1 card move difference (Figure 5 and Table 3). This suggests that the explicit biological framing and requirement to sort into the researcher’s four deep feature categories in the framed task
condition appeared to slightly sharpen but not fundamentally alter the biology faculty population’s performance on this task.

In contrast, the framed task condition did not appear to primarily shift non–biology majors’ results toward a more expert framework based on deep features. As a population, non–biology majors did, however, exhibit a dramatic reduction by 24.6% in the proportion of surface feature card pairings they constructed, with a larger increase in the proportion of unexpected card pairings (14.2% increase) compared with the increase in deep feature card pairings (10.4% increase; Figure 4 and Table 2). This suggests that non–biology majors may have recognized that a surface feature approach was not possible in the framed task condition, but they were unable to make use of the given deep feature group names. Similarly, examination of the edit distance metric shows that non–biology majors shifted away from the hypothesized surface feature sort, and by definition, then, toward the hypothesized deep feature sort. However, this shift was from an ~8 card move difference from a deep sort feature sort in the unframed task condition to an ~6 card move difference from a hypothesized deep feature sort in the framed task condition (Figure 5 and Table 3). Taken together with the data on shifts in proportion of card pairings, these data suggest that non–biology major performance in the framed task condition may reflect their abandonment of the surface feature organizational framework as opposed to an active shift toward more deep feature and expert-like sorting.

Both Non–Biology Majors and Biology Faculty Generate Unexpected Card Pairings. Both participant populations produced unexpected card pairings, namely card pairings that neither represented a hypothesized surface feature pair (e.g., NL, two human cards; Figure 1) nor a hypothesized deep feature pair (e.g., KH, two evolution cards; Figure 1). The proportion of unexpected card pairings was always lower for biology faculty than non–biology majors (Figure 4 and Table 2). However, it is intriguing that the proportion of unexpected card pairings increased between the unframed and the framed task condition for non–biology majors (Figure 4 and Table 2). One possible explanation of this result is that, as described above, the framed task condition and the explicitly given deep feature categories were mostly inaccessible and not understood by the non–biology majors, in contrast to biology faculty. This would lead to the conclusion that the non–biology majors population as a whole did not organize their biological ideas with respect to fundamental biological ideas and that they were unable to recognize and use the organizational framework of fundamental biological ideas given to them in the framed task condition. This also suggests that this particular population of undergraduate students did not have a partial expert framework that could be revealed by explicitly showing them deep feature category names in the framed task condition.

Another key observation was that biology faculty and non–biology majors produced different types of unexpected card pairings. Biology faculty commonly produced unexpected card pairings that were the result of pairing of cards associated with the two particular deep features: “Evolution by Natural Selection in Living Systems” and “Storage and Passage of Information about How to Build Living Systems” (unpublished data). While pairing of cards from these two different deep feature categories would be characterized as an unexpected pair in our analyses, these ideas are intimately connected in biology, with the inheritance of genetic information being the substrate upon which evolutionary mechanisms act.

In contrast, the unexpected pairings produced by non–biology majors were less predictable. No unexpected card pairings were particularly more prevalent than any other. Sometimes these unexpected card pairings appeared to reflect a known misconception. One such example that was seen was the grouping of a card about fungi together with several plant cards (unpublished data). Further detailed analysis of the unexpected card pairings produced by non–biology majors may reveal more patterns.

Unique Aspects of the Biology Card Sorting Task Compared with Other Biology Conceptual Assessment Tools

The Biology Card Sorting Task presented here is intended to expand the repertoire of biological assessment tools available both to researchers from multiple disciplines and to practitioners ranging from individual instructors to departments and larger collaborative initiatives. The Biology Card Sorting Task is unique compared with other currently available biology assessment tools in at least two key ways.

Probing Connections between Conceptual Ideas Rather Than the Ideas Themselves. The Biology Card Sorting Task was designed to assess conceptual expertise in biology—namely how individuals organize their knowledge of biology and how they connect or do not connect biological ideas. While a variety of tools to assess biological conceptual knowledge have been developed, there does not appear to be a tool currently available that probes how individuals organize their biological knowledge broadly across multiple biological ideas. Many biology assessment tools have been developed that probe an individual’s particular knowledge of a specific biological concept such as evolution (e.g., Anderson et al., 2002; Baum et al., 2005, Nehm and Schonfeld, 2008), molecular biology and genetics (e.g., Bowling et al., 2008; Smith et al., 2008; Shi et al., 2010), host–pathogen interactions (e.g., Marbach-Ad et al., 2009), osmosis and diffusion (e.g., Odom and Barrow, 1995; Fisher et al., 2011), and energy and matter in living systems (e.g., Wilson et al., 2007; Hartley et al., 2011), to name just a few. Assessment tools such as these have been designed primarily to measure the presence or absence of particular pieces of knowledge or particular misconceptions about that knowledge, rather than the overall structure and organization of an individual’s biological knowledge. More recently, some researchers have begun to employ these and other tools to explore connections students are making between ideas in a particular concept in biology (e.g., Wilson et al., 2007; Nehm and Schonfeld, 2008; Hartley et al., 2011). The Biology Card Sorting Task may, however, be one of the first tools developed specifically to assess an individual’s conceptual expertise in biology, namely the organization of his or her biological understanding, across a wide range of conceptual ideas in biology.
Assessing How Individuals Will Perform on a Task versus Selection of an Answer Choice. The Biology Card Sorting Task presented here is unique in how an individual’s thinking about biology is probed. Few biology concept assessment tools currently in use are open-ended, requiring individuals to produce evidence of their thinking through either writing or speaking (e.g., Wilson et al., 2007). The majority of currently available assessment tools are closed-ended in their format, and students are asked to select their answer preference from a list of four to five given answers (e.g., Klymkowsky et al., 2003; Sensar et al., 2011). As such, while these assessment tools provide insight into some aspects of student thinking, many other aspects of student thinking are likely going undetected. Dissociations between students’ success in selecting a scientifically accurate answer and an independent analysis of their ability to provide a scientific explanation for a scientifically accurate answer and an independent analysis of some aspects of student thinking and conceptual organization that are not currently being monitored or explored.

Intentionally Unique Aspects of the Biology Card Sorting Task Compared with Other Cognitive Psychology Card Sorting Tasks

While the novel Biology Card Sorting Task presented here was inspired by published card sorting tasks from the developmental and cognitive psychology research literatures, it was intentionally designed to differ from and improve upon previously developed card sorting tasks in four key ways. These design differences were driven by previously described limitations of card sorting tasks generally, as well as by the desire to develop a task-based assessment tool that would be adaptable and feasible for broad use.

**Hypothesis-Driven Card Stimulus Set.** First, the Biology Card Sorting Task was conceptually hypothesis-driven in its design. Previously described card sorting tasks have either not explicitly described why particular card stimuli were selected for use or implied that card stimuli were somewhat randomly selected. In contrast, this Biology Card Sorting Task and its card stimulus set were specifically designed to test a particular hypothesis about how novices might organize their ideas in biology—in terms of organism type (see row identifiers in Figure 1)—and a specific hypothesis about how experts might organize their ideas in biology—in terms of fundamental biological concepts (see column titles of Figure 1). The purposeful development of a hypothesis-driven card set was nontrivial, requiring each card to be able to be seen in the context of both of these two hypothesized conceptual frameworks, as well as to be in language that was as accessible as possible to both novices and experts. Importantly, this approach to constructing a card stimulus set for the Biology Card Sorting Task is now ripe for adaptation and the development of new card stimulus sets that could test different hypotheses about how individuals organize their biological knowledge.

**Probing Individuals’ Own Conceptual Frameworks, as Well as Their Interpretation of Expert Frameworks: Using Two Card Sort Task Conditions with and without Biological Framing.** The Biology Card Sorting Task was also designed to allow individuals both to share their own initial approach to organizing biological ideas (unframed task condition) and to attempt to navigate one possible biological expert framework (framed task condition). Previous card sorting tasks were primarily open-ended, with participants sorting cards into as many groups as they liked (Chi, 2006a,b). In some cases, participants were allowed to do multiple, open-ended sorts with the same card set, referred to as serial repetitive card sorting. However, these tasks did not explicitly probe whether individuals might be able to use an expert framework if explicitly cued to do so.

Because the development of expertise in any discipline is complex, it is no doubt a process that may involve stages. As such, these two (or eventually more) task conditions may be particularly important when studying emerging experts, such as biology majors, and identifying a point at which they may be unable to construct a hypothesized expert framework but may be able to recognize and apply an expert framework. For example, individuals with little to no expertise in biology may be able to come up with some organizational framework for the cards in the unframed task, but then are not able to make sense of an expert framework presented to them in the framed task. In contrast, there may also be individuals with incomplete or intermediate conceptual expertise, such as undergraduate biology majors who may not initially apply an expert framework on their own in the unframed task, but who would be able to navigate and apply an expert framework if explicitly presented with it. To our knowledge, this Biology Card Sorting Task is unique in having the potential to be able to discern putative stages of the development of biological conceptual expertise by assessing both students’ initial sorting strategies and then their actions in response to being given an expert biological framework.

**Multiple Approaches to Quantitative and Qualitative Analysis of Card Sorts.** The data sets generated by conducting the Biology Card Sorting Task are large and complex, including the card groups themselves, the names for the card groups, and the individuals’ narrative responses about their sorting strategies for the two task conditions. Previously published card sorting studies have generally reported results using a nonsystematic and nonquantitative case study method in which an individual’s approach is described and presented (Chi et al., 1981). In a few cases, investigators have quantified some aspects of card sorting results, such as the percent of individuals who assigned a particular card to a particular category, but still with small numbers of participants (reviewed in Chi, 2006a; Mason and Singh, 2011). Given our desire to make this assessment tool useful with large groups of students, we have also developed companion analytical techniques that enable automated, quantitative analysis of card groupings, as well as rubrics to enable blind scoring of the qualitative data that are generated as card group names and rationales for sorting strategies. As such, data entry of card groups generated by an individual is automatically analyzed via a Python computer script to generate quantitative metrics—percent hypothesized surface feature card pairs, percent deep feature card pairs, percent unexpected card pairs, edit distance
from a hypothesized novice sort, and edit distance from a hypothesized expert sort—for each individual’s produced card sorts in both the unframed and the framed task conditions. Analysis of the qualitative data generated from card group names and rationales for sorting strategies is not currently automated; however, rubrics exist for coding of these data. The Biology Card Sorting Task may afford investigators the opportunity to be both systematic and quantitative in analyzing the rich data set that emerges from conducting these assessment tasks with large numbers of individuals.

**A Tool for Classroom Assessment, Program Assessment, and Biology Education Research.** The Biology Card Sorting Task presented here offers the benefit of using a complex, open-ended task to assess the development of biology conceptual expertise that could be used in both research laboratories and classrooms. One aspiration for the Biology Card Sorting Task was to develop a novel task, grounded in the theoretical frameworks of cognitive psychology, that could also be used by biology departments with large numbers of students to provide insight into the progression of student conceptual thinking in biology over the course of an undergraduate degree program. Previous card sorting tasks and protocols that have been developed to study expertise have largely been designed for individual, think-aloud interview protocols, which by nature limit their use to research laboratories. In addition, it is often argued that concept inventories are chosen as an assessment tool due to the ease of quickly analyzing data that emerges from such a closed-ended tool. While many acknowledge that open-ended assessment tools capture a richer profile of student thinking, the resulting data are simply too complex and time-consuming to analyze. The Biology Card Sorting Task is an attempt to develop an approach to measuring the development of discipline-based conceptual expertise that generates a rich profile of student thinking on a complex task, while also generating data that are feasible to analyze at the level of classroom assessment, program assessment, and in the context of biology education research.

**Characteristics of This Card Stimulus Set and Potential Adaptations**

The Biology Card Sorting Task presented here is a first attempt to translate a promising measurement methodology from the cognitive sciences into an assessment tool that can be used in biology education and research. The card stimulus set presented here, however, is far from the only possible card stimulus set that could be used in measuring the development of biological expertise. In addition, the general structure of this card sorting, task-based assessment could be useful in measuring a variety of aspects of biological expertise in a variety of subdomains of this discipline, as well as across science disciplines.

The card stimulus set presented here was designed to test the specific hypothesis that putative biological novices would connect and organize biological ideas using an organizational framework consisting of four specific deep features: evolution, structure–function, information flow, and transformations of energy and matter. While these four core ideas align rather well with recently published frameworks attempting to delineate core ideas in the biological sciences (AAAS, 2011; College Board, 2013), the alignment is by no means exact. For example, the *Vision and Change* document (AAAS, 2011, p. 12), proposes five core concepts for biological literacy that consist of the four deep features we have used here and an additional core concept entitled *systems*. Similarly, the revised organizational framework for AP Biology recently put forward by the College Board includes four core concepts: evolution, genetics and information transfer, cellular processes: energy and communication, and interactions (College Board, 2013). Two of these core ideas align well with two deep features of the Biology Card Sorting Task (e.g., evolution, and genetics and information transfer); the third core idea—cellular processes: energy and communication—may be somewhat more expansive than our transformations of energy and matter deep feature. The core idea of interactions is similar to the systems core idea in *Vision and Change* and could be a separate deep feature in a future card stimulus set. These comparisons are relevant, because the currently used card stimulus set may need further refinement to be maximally useful in some educational settings. Our own results suggest that the addition of a systems or interaction deep feature category to the current card stimulus set would bring the Biology Card Sorting Task into better alignment with some policy documents (AAAS, 2011).

Additionally, this initial card stimulus set was designed to test the hypothesis that putative biological novices would connect and organize biological ideas using the framework of four specific surface features based on organism type: “Human,” “Plant,” “Insect,” and “Microorganism.” While this hypothesized novice framework for organizing biological knowledge appeared to capture ~40% (Figure 7 and Table 5) of the non–biology majors approaches to sorting, a majority of the non–biology majors appear to be sorting using organizational frameworks that we are not yet able to characterize. Adaptation of the current card stimulus set to test other hypothetical frameworks that novices may be using to organize their biological knowledge would produce another variation of the Biology Card Sorting Task.

Finally, multiple agencies are calling for more interdisciplinary courses and programs in undergraduate science education, yet there are few assessment tools available to investigate the extent to which students are connecting knowledge across the disciplines (AAMC–HHMI, 2009). An adaptation of this Biology Card Sorting Task could produce a task-based assessment tool to probe whether students in interdisciplinary courses and programs are more likely than students in traditional courses to sort problems into surface features, such as chemistry, physics, and biology, versus deep features, such as energy and matter and structure and function. The novel structure and quantitative analytical approaches of the Biology Card Sorting Task may be useful to discipline-based science education researchers across the sciences to test an infinite number of hypotheses about how a variety of different populations organize their conceptual knowledge within and across the scientific disciplines.

**CONCLUSIONS**

In conclusion, we have developed a novel assessment tool for the biological sciences that moves away from assessing
individual pieces of knowledge and moves toward measuring how individuals organize biological ideas and develop biological conceptual expertise. This initial investigation of the Biology Card Sorting Task demonstrated that biology faculty and non–biology majors differed significantly on each of six analytical measures used to compare the card sorts produced by these two groups. As such, these data confirm that this novel assessment task appears to be robust in distinguishing populations of putative biology experts and putative biology novices on a variety of measures. With this shown, investigations of how individuals with various levels of biology experiences perform on the Biology Card Sorting Task may now be conducted. Finally, the general structure of the Biology Card Sorting Task may be adaptable for use in assessing other aspects of developing expertise in the biological sciences and beyond.

REFERENCES


Understanding Clicker Discussions: Student Reasoning and the Impact of Instructional Cues

Jennifer K. Knight,* Sarah B. Wise,† and Katelyn M. Southard‡

*Department of Molecular, Cellular, and Developmental Biology and †Department of Ecology and Evolutionary Biology, University of Colorado Boulder, Boulder, CO 80309; ‡Department of Molecular and Cellular Biology, University of Arizona, Tucson, AZ 85721

Submitted May 6, 2013; Revised June 11, 2013; Accepted July 2, 2013
Monitoring Editor: Clarissa Ann Dirks

Previous research has shown that undergraduate science students learn from peer discussions of in-class clicker questions. However, the features that characterize such discussions are largely unknown, as are the instructional factors that may lead students into productive discussions. To explore these questions, we recorded and transcribed 83 discussions among groups of students discussing 34 different clicker questions in an upper-level developmental biology class. Discussion transcripts were analyzed for features such as making claims, questioning, and explaining reasoning. In addition, transcripts were categorized by the quality of reasoning students used and for performance features, such as percent correct on initial vote, percent correct on revote, and normalized learning change. We found that the majority of student discussions included exchanges of reasoning that used evidence and that many such exchanges resulted in students achieving the correct answer. Students also had discussions in which ideas were exchanged, but the correct answer not achieved. Importantly, instructor prompts that asked students to use reasoning resulted in significantly more discussions containing reasoning connected to evidence than without such prompts. Overall, these results suggest that these upper-level biology students readily employ reasoning in their discussions and are positively influenced by instructor cues.

INTRODUCTION

Many lines of research have shown that a wide variety of in-class active-learning activities designed to engage students through discussion of conceptually challenging questions lead to improved student outcomes in undergraduate biology classes (e.g., Udovic et al., 2002; Kitchen et al., 2003; Freeman et al., 2007; Walker et al., 2008). Studies of the use of clickers, remote response systems that students use to answer questions in class, have found that discussion among students (peer discussion) increases performance on these clicker questions (Mazur, 1997). Even without instructor input, student performance improves when students individually answer similar but new questions on a topic. This improvement occurs even when few students initially know the correct answer (Smith et al., 2009) and is independent of overall achievement level of the individual students (Smith et al., 2011). When instructors embrace such peer discussion and establish “sense-making” norms, for example, getting students to understand and articulate their reasoning, this behavior also impacts student attitudes about their learning experience. In such classes, students place a higher value on articulating their reasoning than students in classrooms in which the instructor focuses on getting correct answers (Turpen and Finkelstein, 2010). Thus, there is evidence to support both the idea that students are capable of constructing their own knowledge through peer discussion and that certain instructional practices positively affect student attitudes toward discussion.

Nonetheless, instructors who otherwise embrace using in-class activities sometimes hesitate to encourage small-group discussion due to various concerns, including losing classroom control, having discussions take too much time...
(requiring a reduction in content), and skepticism about whether students are capable of teaching themselves (Felder and Brent, 1996). Therefore, there remains a need to provide evidence that specific instructional methods can lead students into using in-class discussion time productively and that these discussions are more beneficial to students’ development of skills and knowledge than lecturing.

Student discussions have been studied extensively in the K–12 classroom environment. Much of this work has focused on how students exchange ideas, disagree with one another, or support their ideas with reasons. In general, these interactions are referred to as “argumentation.” Toulmin (1958) defined quality argumentation as a situation in which the speaker makes a claim, provides evidence or reasoning for his/her claim, and ultimately provides a warrant—a statement that links the initial claim to the supporting evidence. Toulmin’s classification of the different elements of argumentation has been subsequently used and modified by others to help describe and characterize student dialogue (Driver et al., 2000; Jiménez-Aleixandre et al., 2000; Sampson and Clark, 2008). Some have focused on the correct use of scientific content (Sampson and Clark, 2008) or the frequency of rebuttals, in which a student challenges another student’s initial offering of evidence (Osborne et al., 2004), while others have explored how teachers’ questions and prompts impact the nature of the discussion (Michaels et al., 2002).

In terms of fostering student argumentation in the classroom, two conditions appear particularly important: an instructional task that challenges students to consider alternate ideas and a social context that invites dialogue (Osborne et al., 2004). Many secondary school science classrooms do not provide opportunities to engage in argumentation (Lemke, 1990), and, when students are given the opportunity, they often do not readily employ reasoning in their arguments (Kuhn, 1991; Kuhn and Udell, 2003; Zohar and Nemet, 2002). However, when instructors use such behaviors as prompting students to use reasoning or modeling what such reasoning should look like, student argumentation skills of even young students improve, suggesting that argumentation is a skill that must be explicitly taught and practiced (Zohar and Nemet, 2002; Osborne et al., 2004; McNeill et al., 2006).

In contrast to the wealth of information on younger students, few studies have examined the content or nature of argumentation in college-level classrooms. When asked to construct a written argument to explain data, undergraduates in an introductory biology class were able to generate simple features of making a claim and using evidence but did not generally supply warrants for their reasoning or construct rebuttals unless explicitly directed to do so (Schen, 2012). They also struggled with providing alternative explanations for data, even when prompted. Similar patterns have been found in the analysis of oral argumentation. In a large introductory astronomy class, less than half of student clicker question discussions involved an exchange of claims and support or rebuttals of those claims with additional discussion (James and Willoughby, 2011). The majority of discussions involved considering ideas not presented in the clicker question or answers, showed a lack of understanding of basic principles needed to discuss the question, or went off task. In addition, some “discussions” mainly involved group members listening to the dominant person in the group rather than exchanging reasoning, especially when correct answers were rewarded with more points (also previously discussed in James, 2006). These studies suggest that college-level students are similar to younger students in their lack of argumentation skills and should benefit from explicit instruction in how to use discussion time productively.

If faculty members could be confident that certain techniques would lead their students to engage in meaningful discussions that improve learning, they might be more likely to implement such techniques. Thus, it is important both to further characterize the content of peer discussions in different types of undergraduate science classes and to identify ways that instructors can successfully influence the quality of student discussion. In this study, we have characterized the types of statements upper-level undergraduate biology students make when engaged in discussions of challenging conceptual clicker questions and have explored the relationship of these statements to the outcome of the discussions and whether explicit instructor cues affect the amount and kind of reasoning students use when given the opportunity to discuss.

METHODS

Characterization of the Course, Students, and Instructional Style

The students in this study were enrolled in an upper-division developmental biology course, one of several possible required capstone courses taken by majors in their junior or senior year. The course (taught by J.K.K.) was designed to challenge students to apply knowledge gained in four prerequisite core courses and to provide practice in critical thinking and problem solving. The course met for 75 min twice a week for 15 wk. The 107 students enrolled in the course (47% female, 97% majors) were seated at 12 round tables with 8–10 students per table. Students were allowed to self-select into groups, which were formed in the first week of class with encouragement from the instructor and maintained for the rest of the semester. The class was taught in an active, student-centered style: < 60% of class time was spent lecturing; students completed written in-class problem-solving exercises in at least 70% of class periods; and students were asked an average of five clicker questions per class period. Clicker questions were implemented following the peer instruction model (Mazur, 1997): students answered questions initially on their own; the instructor encouraged peer discussion (without showing the vote distribution) when less than 70% of students initially answered the question correctly; and students revoted after discussion. Following discussion and revote, volunteers were usually asked to share their ideas and problem-solving strategies with the rest of the class, and the instructor wrapped up discussion after showing the histogram distribution of the revote.

Instructor Cues

During the semester, the instructor varied how clicker questions were introduced prior to and following peer discussions, taking either an “answer-centered” or “reasoning-centered” approach (Table 1). The instructor alternated between these two kinds of cues on a weekly basis throughout the semester: for example, in one week, the instructor used...
Instructor wrap-up  Instructor explained reasoning and correct answer.

Histogram of student answers  Shown immediately after discussion and revote.

Instructor highlighted student explanations that described correct answer.

Table 1. Instructional behaviors in answer-centered and reasoning-centered class periods

<table>
<thead>
<tr>
<th></th>
<th>Answer-centered class</th>
<th>Reasoning-centered class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instructor cue</td>
<td>“Discuss your answers with your table and focus on the reasons for your answers. Then, I’ll ask you to share your reasons.”</td>
<td>“Discuss your answers with your table and focus on the reasons for your answers. Then, I’ll explain the correct answer.”</td>
</tr>
<tr>
<td>Student reasoning requested</td>
<td>None</td>
<td>Volunteers were asked to share reasoning from their group’s discussion.</td>
</tr>
<tr>
<td>Instructor wrap-up</td>
<td>Shown immediately after discussion and revote Instructor explained reasoning and correct answer.</td>
<td>Shown after students volunteered reasoning.</td>
</tr>
</tbody>
</table>

Due to the nature of the recording equipment, discussions sometimes involved inaudible portions of conversation. However, recordings typically included a logical thread of conversation among two to five students. Discussions that were mostly inaudible were not included in this study.

Analysis: Fine-Grained Discussion Coding

We identified each student statement as a turn of talk—if a student spoke several sentences in a row, this was still defined as one turn of talk. We then gave each turn within a transcript one or more codes representing its role in the discussion. We tracked several elements of discussion: claim making (stating their choice of an answer), explaining reasoning for an answer, asking a question, or providing background information (Table 2). Statements that were off topic or simple statements of agreement or disagreement (“Okay,” or “That doesn’t make sense”) were given a code of NA (not applicable). Turns of talk were given more than one code if they included more than one element (e.g., a statement might have been coded as both a question and a background statement), or if they included more than one instance of the same element (e.g., multiple pieces of evidence used to support a reason). The coding scheme was determined and refined by four coders; after an interrater agreement of greater than 0.7 on 10% of the transcripts was achieved, the remaining transcripts were independently coded. The total turns of talk were calculated for each discussion, as was the fraction of the discussion devoted to each of the codes described above.

Analysis: Whole-Discussion Measures

In addition to fine-grained coding, each transcript was categorized on a whole-discussion level for several different features. After considering other whole-discussion coding

Table 2. Definition of fine-grained codes

<table>
<thead>
<tr>
<th>Code</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Claim</td>
<td>A statement of preference for an answer (such as “I think it’s ‘A’”)</td>
</tr>
<tr>
<td>Reasoning</td>
<td>A unique explanation for choosing or eliminating an answer, including warrants and rephrasings of previously given reasons</td>
</tr>
<tr>
<td>Question</td>
<td>Any question (asking peer to explain idea, asking about definitions, wording, or background information)</td>
</tr>
<tr>
<td>Background</td>
<td>Providing information about a question to clarify what the question was asking or what a figure showed</td>
</tr>
<tr>
<td>NA</td>
<td>Any speech not described by the above codes, including simple statements of agreement or disagreement, reasoning statements repeated verbatim, joking, and off-topic talk</td>
</tr>
</tbody>
</table>

*Each turn of talk was given one or more codes as defined.*
schemes (e.g., Osborne et al., 2004), we developed a scoring system to measure the quality of student reasoning (Exchange of Quality Reasoning; Table 3) based on exchanges of warrants (Toulmin, 1958). The highest-scoring discussions, level 3, feature multiple students linking evidence to a claim with logical reasoning; while the lowest-scoring discussions involve no exchange of reasoning although they usually involve exchanges of claims. Each discussion was scored independently by J.K.K. and S.B.W. (interrater agreement of 0.86); any differences were resolved by consensus. We also tracked discussions containing conflicting lines of reasoning, in which one student’s reason was challenged by at least one other student’s reason.

Several additional whole-discussion measures were taken: the amount of time spent in discussion, the percent correct on the initial and revote for all students at each recorded table, and the normalized change in voting for each recorded table of students discussing a clicker question. The normalized change formula uses the original normalized gain formula (Hake, 1998) of: 100(revote − initial)/(100 − initial) for positive changes, and the formula: 100(revote − initial)/initial for negative changes (Marx and Cummings, 2007). If the percent correct was 100% for both initial and revote, the discussion was not included in the average calculation.

## RESULTS

### Characterizing Discussions

#### General Characterization of Clicker Questions and Clicker Discussions

Over the semester, the instructor posed a total of 124 clicker questions. When the initial vote was < 70% correct, the instructor asked students to discuss their answers and revote: 81 questions fell into this category. Each discussion lasted an average of 2 min (± 0.93 SD) and involved from two to five students. During each class in which recordings were made, one to three clicker questions were selected at the end of each class period by the instructor for transcription, based on the clarity of the questions (one correct answer) and their likelihood of generating discussion. After listening to these recordings, we had to exclude some due to poor audio quality. Ultimately, discussions of 34 questions from the 17 class periods in which recordings were made were transcribed and further analyzed. Multiple groups were recorded discussing each question, resulting in a total of 83 transcripts. In most cases, the student votes at a particular table were similar to the class-wide average of < 70% correct. In 17 recorded discussions, the initial vote for students at one or more tables was higher than 70%; these discussions were still transcribed and analyzed.

The 34 clicker questions selected for analysis were rated as either requiring Bloom’s lower-order cognitive skills (LOCS) or higher-order cognitive skills (HOCS; Crowe et al., 2008) by two independent raters and the instructor; a final rating for each question was reached by consensus. Twenty-nine of the questions were rated as requiring HOCS, and five as requiring LOCS.

The 34 transcribed clicker questions were on average more difficult than the complete set of clicker questions for the course (48% correct vs. 61% correct, respectively), because discussions and revotes were only performed when the overall vote was < 70% correct. The students who volunteered to have their discussions recorded did not perform significantly differently from the rest of the students in the class on clicker questions (Figure 1) or on other measures, such as exam performance, suggesting that the volunteers were representative of the class. After peer discussion, the average percent correct on the revote was significantly higher than the average initial vote for both recorded and unrecorded students, as has been reported for other studies on peer discussion (e.g., Smith et al., 2009, 2011).

#### Transcript Coding

**Fine-Grained Coding.** For each transcript, each student turn of talk was characterized using the codes shown and explained in Table 2. The total turns of talk were summed for each discussion and were used to characterize the percent of turns of talk devoted to each feature. On average, students spent a larger percent of each discussion explaining their reasons to each other than making claims, asking questions, providing background, or making NA remarks (Table 4), although there was considerable variability among discussions.

To determine whether the features of each discussion were affected by the initial percent of students who answered each clicker question correctly, we looked for correlations between the initial percent correct and the percent of turns of talk spent

<table>
<thead>
<tr>
<th>Level</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>No students made reasoning statements.</td>
</tr>
<tr>
<td>1</td>
<td>Only one student used reasoning, which could include a warrant (no exchange).</td>
</tr>
<tr>
<td>2</td>
<td>Two or more students exchanged reasoning, but neither or only one included warrants.</td>
</tr>
<tr>
<td>3</td>
<td>Two or more students exchanged reasoning, including warrants.</td>
</tr>
</tbody>
</table>

*Each transcript was assigned a level based on the characteristics described.

<table>
<thead>
<tr>
<th>Level Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 No students made reasoning statements.</td>
</tr>
<tr>
<td>1 Only one student used reasoning, which could include a warrant (no exchange).</td>
</tr>
<tr>
<td>2 Two or more students exchanged reasoning, but neither or only one included warrants.</td>
</tr>
<tr>
<td>3 Two or more students exchanged reasoning, including warrants.</td>
</tr>
</tbody>
</table>

*Each transcript was assigned a level based on the characteristics described.

---

### Table 3. Exchange of Quality Reasoning levels*

<table>
<thead>
<tr>
<th>Level</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>No students made reasoning statements.</td>
</tr>
<tr>
<td>1</td>
<td>Only one student used reasoning, which could include a warrant (no exchange).</td>
</tr>
<tr>
<td>2</td>
<td>Two or more students exchanged reasoning, but neither or only one included warrants.</td>
</tr>
<tr>
<td>3</td>
<td>Two or more students exchanged reasoning, including warrants.</td>
</tr>
</tbody>
</table>

*Each transcript was assigned a level based on the characteristics described.

---

### Table 3. Exchange of Quality Reasoning levels

<table>
<thead>
<tr>
<th>Level</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>No students made reasoning statements.</td>
</tr>
<tr>
<td>1</td>
<td>Only one student used reasoning, which could include a warrant (no exchange).</td>
</tr>
<tr>
<td>2</td>
<td>Two or more students exchanged reasoning, but neither or only one included warrants.</td>
</tr>
<tr>
<td>3</td>
<td>Two or more students exchanged reasoning, including warrants.</td>
</tr>
</tbody>
</table>

---

### Figure 1. Average percent correct on transcribed clicker questions for unrecorded (average n = 60) and recorded (average n = 35) students. The initial and revote values are not significantly different between the two groups (p > 0.05; two-way repeated measures ANOVA). Revote values are significantly higher than initial for both groups (p < 0.001; two-way repeated measures ANOVA, no interaction p = 0.88). Error bars indicate SEM.
on each of the fine-grained coding features (Tables 2 and 4). No significant correlations exist (Pearson’s r values: all $p > 0.05$). For example, in transcripts for which students began with a low initial percent correct, some discussions devoted a high percentage of the discussion to reasoning, and others a low percentage. In the 17 discussions for which the initial vote for that group of students was above 70% correct, the students still spent on average 42% of the discussion explaining their reasoning; similar to the overall average for all discussions (39%).

To determine whether the percent of turns of talk spent on these features subsequently impacted the percent of students who answered correctly on the revote, we also looked for correlations between these two measures. As was true for the initial vote, no correlations exist between individual features and the percent correct on revote (Pearson’s $r$ values: all $p > 0.05$). Discussions in which a high proportion of the discussion was spent exchanging reasoning or a high proportion was spent exchanging claims were no more likely to produce a high correct revote than were discussions in which a low proportion of time was spent on reasoning or claims. Thus, the fraction of discussion spent on different features of talk is not necessarily predictive of the vote outcome.

Finally, there was no correlation between time of the semester and each of the measures reported above; although individual discussions vary in student use of reasoning and in student performance, there are no positive or negative trends from beginning to end of the semester (Pearson’s $r$ values: all $p > 0.05$).

**Whole-Discussion Coding of Reasoning.** Reasoning statements for all discussions were coded as described above. Some reasoning statements were consensus building, in which a student agreed with a previous student, and added to the line of reasoning. Some reasoning statements were conflict oriented, in which a student disagreed with a line of reasoning previously made and gave a conflicting reason to support his or her claim. Willingness to engage in disagreements with one another was quite common: 66% of the discussions included conflicting reasoning statements.

To characterize the reasons that students provided, we scored each of the discussions using the Exchange of Quality Reasoning measure (Table 3). This measure emphasizes the value both of exchanging reasoning and of providing reasoning that logically connects evidence with a claim (warrants).

An example of two discussions that illustrate the difference between a level 1 and a level 3 score for Exchange of Quality Reasoning is shown in Figure 2. The discussions are about the same clicker question, and both began with a low percent correct initial vote (15–18%). In discussion A, students used many reasoning statements to support their claims, and their reasoning statements included warrants connecting their reasons with evidence. In discussion B, only one student gave a reasoning statement, while the rest of the students asked questions and made background comments. Discussion A was scored a 3 and resulted in 85% of the students at the table answering the question correctly in the revote; discussion B was scored a 1, and 0% of the students at the table answered the question correctly in the revote. Most of the 83 discussions were categorized into the two highest-quality reasoning levels, both of which involve exchanges of reasoning (level 3: 54%; level 2: 24%). A smaller number involved reasoning provided by one student only (level 1: 18%), and only three discussions exhibited no reasoning at all (level 0: 4%). Because so few discussions scored “0” level, we combined discussions from levels 0 and 1 for additional analysis.

To examine whether discussions scored at different levels of reasoning resulted in different learning outcomes, we compared the mean percent correct on revotes and the normalized change in correct voting that followed discussions of each level (Figure 3). Discussions that involved an exchange of warrants (level 3) had the highest normalized change and the highest percent correct on revotes. However, correct revoting was not significantly different between level 3 and the other levels of reasoning. As it is not valid to apply statistical tests to normalized change values (Marx and Cummings, 2007), only the SE bars for each value are displayed in Figure 3.

To assess whether any fine-grain coded features reported above are associated with discussions that score highly for reasoning quality, we tested the relationship between the frequency of a discussion feature and its quality-of reasoning level. Discussions that involved an exchange of reasoning (levels 2 and 3) devoted a significantly higher percent of the discussion to reasoning and also had significantly more turns of talk, compared with level 0/1 discussions (Table 5). In other words, when students exchanged ideas, they engaged in longer discussions and provided a significantly greater number of statements of reasoning than in scenarios in which only one person was offering reasons for his or her answer.

Finally, we investigated whether there was a correlation between Bloom’s level of a question and the quality of reasoning that students used when discussing the question. The discussions in the 0/1 quality of reasoning level were distributed over 15 of the 34 clicker questions. Three of the 15 questions were judged as requiring LOC- rather than HOC-level skills and resulted in five of the 18 level 0/1 discussions, suggesting that these particular questions were less rich for generating discussion than other questions in the data set. However, the two other LOC clicker questions generated level 2/3 discussions in all recorded groups.

**Differences between Groups of Students.** We recorded from four different tables of students over the course of the semester. One table chose to exit the study after a few weeks; we therefore recruited another table to take its place. Thus, the number of transcripts obtained was not equal among the tables of students, and it was necessary to investigate whether table identity affected the results. We found that tables did not differ significantly in the distribution of quality-of-reasoning

| Table 4. Prevalence of each code within clicker discussionsa |
|-----------------|-----------------|
| Code            | Average percentage of turns of talk for all discussions ($n = 83$) |
| Claim           | 30.8 (1.9)      |
| Reasoning       | 39.2 (2.4)      |
| Question        | 17.7 (1.0)      |
| Background      | 5.7 (1.0)       |
| NA              | 28.3 (1.7)      |

aEach transcript was treated as a single discussion. Discussions had on average 20.5 (1.2) turns of talk. Values do not total 100%, because each turn of talk could be given more than one code. SEM shown in parentheses.
levels in their transcripts (Kruskal Wallis: $p = 0.07$); the fraction of discussions that included conflicting reasoning (chi-squared: $p = 0.55$); or the measures of initial percent correct, revote percent correct, or normalized learning change (one-way ANOVA: $p > 0.05$). There do exist differences in some of the fine-grained coding measures: one table devoted more of their discussions to claims, while a different table spent more of their discussions on reasoning than did the other tables (one-way ANOVA, Tukey post hoc: $p < 0.05$). Because the discussions of each table differed significantly on very few measures, and within-table variability exceeded between-table variability, analyses of discussions reported above were done with all tables aggregated.

Further qualitative analysis of the discussions of different tables revealed some interesting features. One table was especially collaborative: students added to one another’s ideas, collectively building on an often correct but incomplete idea to ultimately create a well-supported reason for their answer. When they did initially disagree on the correct answer, they provided evidence to support their conflicting reasons, exchanging ideas until they arrived at an answer that made sense to the group as a whole. Although not statistically significant, this group had the highest percent correct both on their initial votes (60.5%) and on their revotes (87.5%), suggesting that despite an initially high understanding, they readily engaged in further discussion, helping others at the table to achieve understanding. Another table had a different style of interacting: they made more jokes and were more (good-naturedly) critical of one another. This table made the most claim statements and often disagreed on their initial vote. Although not significantly different from other tables, their discussions frequently contained conflicting pieces of evidence to support their claims. They had a lower initial percent correct than the previous table (54.5%) but still a reasonably high average percent correct on revote (74.5%). Both these tables performed slightly but not significantly better with respect to revote percent correct than the other two tables, which spent more time expressing frustration and uncertainty or listening to one person who appeared to have an idea of the correct answer.

**Impact of Instructional Cues.** To measure the impact of different instructional cues on student discussion, we separated discussions into two categories: those following reasoning-centered cues (51 discussions), and those following answer-centered cues (32 discussions; Table 1). Discussions in the two categories did not differ in the average time spent discussing clicker questions, the average percent correct on initial vote, the average percent correct on revote, or the fraction of the discussion spent on reasoning (Table 6), paralleling our finding that many discussion features are not directly correlated with measures of performance. However, when the instructor
Table 5. Characteristics of discussions scored by Exchange of Quality Reasoning

<table>
<thead>
<tr>
<th>Exchange of Quality Reasoning level</th>
<th>Number of discussions</th>
<th>Average turns of talk per discussion (SEM)</th>
<th>Average percent of discussion devoted to reasoning (SEM)</th>
<th>Average percent correct on revote</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>45</td>
<td>24.7 (1.7)</td>
<td>44 (2.7)</td>
<td>79.8</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>20 (1.6)</td>
<td>43 (6.0)</td>
<td>69.1</td>
</tr>
<tr>
<td>0/1</td>
<td>18</td>
<td>10.7 (1.0)*</td>
<td>22 (3.2)*</td>
<td>70.5</td>
</tr>
</tbody>
</table>

*Level 0/1 discussions significantly lower than levels 2 and 3; *p* < 0.05 (one-way ANOVA). Level-2 and level-3 discussions were not significantly different from each other on any of these measures.

used reasoning cues, students engaged in significantly more high-quality discussions that included exchanges of warrants (level 3) than when the instructor cued students to focus on the answer. In turn, the fraction of the discussion spent on claims was significantly lower in reasoning-cued discussions. Reasoning-cued discussions were also more likely to exhibit conflicting lines of reasoning among students (73%) than were answer-cued discussions (56%), although this difference is not statistically significant (Table 6).

DISCUSSION

In this study we characterized the features of high- and low-quality peer discussions of in-class clicker questions among upper-division undergraduate biology majors. We analyzed how the features of these discussions related to performance, and we discovered that certain features of discussion differ in response to instructor cues.

Upper-Division Students Generally Engage in Productive Discussion

We find that students in general vote more correctly following peer discussion, supporting previous work (Smith et al., 2009, 2011) and indicating that their engagement in peer discussion improved their understanding (Figure 1). In contrast to introductory astronomy students (James and Willoughby, 2011), recorded volunteers in this upper-division course engaged in the type of discussion the instructor intended for almost all of the transcripts analyzed: that is, they exchanged reasoning related to the clicker question asked. In only three cases did students fail to discuss their ideas after exchanging information about their votes.

Smith et al. (2009) suggested that improvement of student performance on clicker questions likely results from a

Table 6. Comparison of answer-cued and reasoning-cued discussions*

<table>
<thead>
<tr>
<th></th>
<th>Answer cued (n = 32)</th>
<th>Reasoning cued (n = 51)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (minutes)</td>
<td>2.5 (0.5)</td>
<td>2.7 (0.8)</td>
</tr>
<tr>
<td>Turns of talk</td>
<td>18.7 (1.4)</td>
<td>21.7 (1.7)</td>
</tr>
<tr>
<td>Percent correct initial vote</td>
<td>56.9 (3.6)</td>
<td>48.8 (3.6)</td>
</tr>
<tr>
<td>Percent correct revote</td>
<td>80.5 (5.3)</td>
<td>72.9 (4.9)</td>
</tr>
<tr>
<td>Percent devoted to claims</td>
<td>36.2 (3.4)*</td>
<td>27.4 (2.1)</td>
</tr>
<tr>
<td>Percent devoted to reasoning</td>
<td>33.4 (3.4)</td>
<td>42.8 (3.2)</td>
</tr>
<tr>
<td>Exchange of Quality Reasoning level</td>
<td>2.0 (0.1)</td>
<td>2.5 (0.1)**</td>
</tr>
<tr>
<td>Percent of discussions that involved conflicting lines of reasoning</td>
<td>56</td>
<td>73</td>
</tr>
</tbody>
</table>

*Average time, turns of talk, and percent correct initial and revote were not significantly different between the two conditions (independent samples *t* test: *p* > 0.05 in all cases). The percent of discussions that involved conflicting lines of reasoning was also not significant (Mann-Whitney *U*-test: *p* = 0.129). All values shown are averages. SEM is shown in parentheses.

*p* < 0.05, statistically significant difference between answer-cued and reasoning-cued discussion (independent samples *t* test).

**p** < 0.01, statistically significant difference between answer-cued and reasoning-cued discussion (Mann-Whitney *U*-test).
cooperative group construction of knowledge (coconstruction), rather than simply from one student telling other students the correct answer (transmission). We found evidence of coconstruction in more than three-quarters of these discussions, which exhibited either an exchange of reasoning (level 2) or an exchange of warrants (level 3). It is possible that this comparatively high-level reasoning behavior may be explained by the students’ upper-division status. These students had used clickers in most of their core biology courses and likely had opportunities to develop sophistication in reasoning, particularly in so-called critical-thinking courses, which involve reading and presenting on original literature. Presumably, most upper-division biology majors would also have developed at least some of these skills through their course work. Finally, the students in this study may have further developed reasoning skills specifically during this course as they became more familiar with answering and discussing challenging clicker questions. If this development occurred, it likely had the largest impact prior to the onset of recording in week 3 of class, because we did not see evidence of an improvement in quality of reasoning over the semester. However, such a change as a result of in-course practices may occur in less advanced students, such as those enrolled in introductory biology, who are the target audience of our future studies.

The Bloom’s Level of a Question Does Not Necessarily Influence Student Discussion

Answering most clicker questions in this study required Bloom’s HOC skills. We found that the five questions rated as requiring LOC skills still had the potential to generate student discussion that involved exchanges of reasoning. This finding is consistent with James and Willoughby’s work (2011), in which the authors reported that introductory astronomy students discuss “recall” (Bloom’s level 1) questions extensively, despite instructors’ perceptions that these questions are simple or basic. Thus, the cognitive level of a question does not necessarily correlate with its perceived easiness or difficulty as judged by instructors (Lemons and Lemons, 2013) and does not determine the quality of the discussion among students.

Initial Votes on Clicker Questions Do Not Determine Discussion Features

In this course, when the class-wide vote was above 70% correct, the instructor did not have students discuss and revote on the question. However, recorded discussions among groups of students varied—sometimes the initial vote at a table was above 70% correct, even though the class-wide vote was below 70%. Thus, we had the opportunity to investigate how students discussed questions when many of them were already in agreement about the correct answer. We found that, when students already had the correct answer, they still discussed their ideas and were just as likely to exchange claims, questions, and warrants as when they did not already have the correct answer. This may suggest that, contrary to previous assumptions, upper-division students find value in pursuing discussion even when many of them already have voted for the correct answer. Some possible reasons for this behavior include: 1) students may initially vote for the correct answer without feeling confident in their explanation for that answer, or 2) students may vote for the correct answer but may have the wrong reasoning to support this answer. Additionally, because this class involved up to 10 students sitting at the same table and working together as a group, it is likely that even a single difference of opinion, if voiced, could spur discussion.

Why Do Students Sometimes Fail to Perform Better after Discussion?

In 25% of the discussions analyzed, individual tables of students did not improve after discussion; in some cases, the percent correct on the revote remained the same, while in other cases, more students selected the incorrect answer than during the individual vote. Even when more than 70% of the class achieved the correct answer after discussion, there were sometimes whole tables of students who did not revote correctly. Reviewing discussions of this nature revealed that this scenario often occurs when one or more students begin with an incorrect idea and are able to use evidence to convincingly support this idea, even though the evidence is not factually correct. In other instances, one student with the correct idea may not supply convincing reasoning or may supply no reasoning at all, in which case, the other students stick with their initial incorrect answer. Finally, in some cases, none of the students have the correct idea, no one is motivated to share his or her reasoning, or no support for the correct idea is offered within a group, leading to no change in the students’ ideas (as evidenced by discussion B in Figure 2). In summary, students who vote incorrectly in this situation are not necessarily isolated individuals spread out in the classroom who did not participate in the discussion. These findings support the recommendation that whole-class discussion of both the most commonly chosen incorrect answer and the correct answer should be helpful for students, even when most students have answered the question correctly (Caldwell, 2007). It also further supports the practice of not showing the histogram of student answers until after students have an opportunity to share their reasoning with the class, so students are not biased in their discussion by the majority vote (Perez et al., 2010).

Instructor Cues Influence Quality of Student Discussion

This course was taught in a student-centered, active style throughout the semester. Class periods were very similar in terms of expectations for student participation and the engagement of students in other in-class activities besides clicker questions. The only appreciable difference in the answer-cued and reasoning-cued class periods was how the clicker question discussions were cued, and how the instructor followed through with class-wide discussion (Table 1). Interestingly, focus group interviews with volunteers who had participated in the recordings revealed that students were not explicitly aware of the different cueing. Thus, even though the cues used involved subtle changes in the patterns of instructor–student interaction, student discussion behavior still shifted significantly in response to differences in these cues.
We propose several possible explanations for the increased quality of reasoning in the reasoning-cued discussions. One possibility is that the students are responding solely to the cue: they hear the suggestion to use reasoning, and this reminder is enough to stimulate such discussion. Another possibility is that students are reacting to the instructor’s cue that they will be held accountable for their discussion: accountability has been shown to impact student reports of their attentiveness to the task at hand (Nichol and Boyle, 2003). In reasoning-cued discussions, students were told they would be asked to explain their tables’ reasons to the rest of the class. This may have motivated students to focus on being able to explain their ideas, thus encouraging them to provide evidence for their claims. Supporting this as a potential mechanism, we noticed several instances in the reasoning-cued discussion of students expressing concern that they be able to explain their ideas: “If she calls on me, I’ll die because I have no idea,” and “But if she asks how do we know we’re right, how do we know it’s the right answer?” Finally, another possibility is that the students are negatively affected by answer-cued instruction. By placing emphasis on achieving a correct answer and by leading students to expect an instructor explanation, students may be prevented from engaging in their “normal” level of discussion. Thus, although accountability may in fact be a strong motivator, further work is necessary to exclude the possibility that answer cueing is demotivating.

Value of Student Discussion

We find it notable that exchanging reasoning does not guarantee that students will arrive at a correct answer: the fraction of each discussion spent explaining reasoning did not correlate with the percent of students at a given table who ultimately answered the question correctly (Table 5). As most discussions contained reasons supporting both correct and incorrect answers, it is not surprising that students sometimes led each other in an incorrect direction. On the other hand, discussions that included exchanges of warrants of those reasons, in which evidence was used to explain a claim or justify an idea, resulted in a higher percent of students answering correctly (Table 5 and Figure 3). Although these were not significant differences, the findings suggest a tendency for the highest-quality discussions to result in more correct answers.

Nevertheless, these trends do not address what we think is a critical benefit of student discussion: the act of discussion itself is an important component of learning, regardless of whether students are immediately able to select the correct answer for a clicker question. Socially mediated communication has been shown to be crucial for an individual’s exposure to and practice with new ideas, and is a frequent prerequisite to individual internalization of concepts (Vygotsky, 1978; Lave and Wenger, 1991). Science content learning is mediated through language and communication, and, in the education process, students have the opportunity to acquire the language of science through discourse (Osborne, 2010). In addition, dialectical argumentation, in which more than one side of an argument is explored, has also been shown to be central in the learning process (Asterhan and Schwarz, 2009). Thus, even when students argue for incorrect answers, they are engaging in this process of learning. In addition, anecdotally, students often refer back to previous discussions of clicker questions. We have observed this behavior both when students are considering new clicker questions and in other problem-solving settings, such as help sessions and homework-solving sessions. This suggests that students are remembering and using previous peer discussions to help them reason through new scenarios.

In summary, we have shown that upper-level biology students in a student-centered course readily discuss their answers to clicker questions by exchanging reasons and providing evidence for their ideas. We have also demonstrated that the initial vote or clicker question type does not determine the amount or quality of the reasoning, suggesting that students can benefit from discussion no matter what the conditions. In addition, we show that students follow numerous paths in their discussion: paths that consider multiple answers and result in a correct answer, as well as paths that lead to an incorrect answer due to exchange of incorrect reasoning, one convincing person with an incorrect idea, or an absence of discussion. Most importantly, the evidence presented in this paper supports a critical role for the instructor in stimulating high-quality discussions of clicker questions. Students changed their discussion behavior in response to instructor cues, using more quality reasoning when the instructor emphasized using and sharing reasoning. Whether these discussions not only help students with the social process of problem solving and understanding material in class, but also impact retention and understanding of concepts long term, deserves further study.

REFERENCES


Understanding Clicker Discussions

Article

The Classroom Observation Protocol for Undergraduate STEM (COPUS): A New Instrument to Characterize University STEM Classroom Practices

Michelle K. Smith,* Francis H. M. Jones,† Sarah L. Gilbert,‡ and Carl E. Wieman‡

*School of Biology and Ecology and Maine Center for Research in STEM Education, University of Maine–Orono, Orono, ME 04469-5751; †Department of Earth, Ocean, and Atmospheric Sciences, University of British Columbia, Vancouver, BC V6T 1Z4, Canada; ‡Carl Wieman Science Education Initiative, University of British Columbia, Vancouver, BC V6T 1Z3, Canada

Submitted August 10, 2013; Revised September 8, 2013; Accepted September 9, 2013
Monitoring Editor: Erin L. Dolan

Instructors and the teaching practices they employ play a critical role in improving student learning in college science, technology, engineering, and mathematics (STEM) courses. Consequently, there is increasing interest in collecting information on the range and frequency of teaching practices at department-wide and institution-wide scales. To help facilitate this process, we present a new classroom observation protocol known as the Classroom Observation Protocol for Undergraduate STEM or COPUS. This protocol allows STEM faculty, after a short 1.5-hour training period, to reliably characterize how faculty and students are spending their time in the classroom. We present the protocol, discuss how it differs from existing classroom observation protocols, and describe the process by which it was developed and validated. We also discuss how the observation data can be used to guide individual and institutional change.

INTRODUCTION

A large and growing body of research indicates that undergraduate students learn more in courses that use active-engagement instructional approaches (Prince, 2004; Knight and Wood, 2005; Michael, 2006; Blanchard et al., 2010). As a result, the importance of teaching science, technology, engineering, and mathematics (STEM) courses more effectively has been stressed in numerous reports, including the President’s Council of Advisors on Science and Technology Engage to Excel report (2012), the National Science Foundation/American Association for the Advancement of Science Vision and Change report (AAAS, 2010), and the National Research Council Discipline-Based Education Research report (Singer et al., 2012). Given these compelling, evidence-based recommendations and the recognized need for measures of teaching effectiveness beyond student evaluations (Association of American Universities, 2011), higher education institutions are struggling to determine the extent to which faculty members are teaching in an interactive manner. This lack of information is a major barrier to transforming instruction and evaluating the success of programs that support such change.

To collect information about the nature of STEM teaching practices as a means to support institutional change, faculty at both the University of British Columbia (UBC) and the University of Maine (UMaine) created classroom observation programs. The results of such observations were needed to: 1) characterize the general state of STEM classroom teaching at both institutions, 2) provide feedback to instructors who desired information about how they and their students were spending time in class, 3) identify faculty professional development needs, and 4) check the accuracy of the faculty reporting on the Teaching Practices Survey that is now in use at UBC (CWSIE Teaching Practices Survey, 2013).

To achieve these goals, the programs needed an observation protocol that could be used by faculty member observers.
to reliably characterize how students and instructors were spending their time in undergraduate STEM classrooms. A critical requirement of the protocol was that observers who were typical STEM faculty members could achieve these results with only 1 or 2 hours of training, as it is unrealistic to expect they would have more time than that available. In the quest for a suitable observation protocol, multiple existing options were considered, and ultimately rejected.

The observation protocols considered were divided into two categories: open-ended or structured. When observers use open-ended protocols, they typically attend class, make notes, and respond to such statements as: “Comment on student involvement and interaction with the instructor” (Millis, 1992). Although responses to these types of questions can provide useful feedback to observers and instructors, the data are observer dependent and cannot easily be standardized or compared across multiple classrooms (e.g., all STEM courses at UBC or UMaine).

Alternatively, structured protocols provide a common set of statements or codes to which the observers respond. Often, these protocols ask observers to make judgments about how well the teaching conforms to a specific standard. Examples of such protocols include the Inside the Classroom: Observation and Analytic Protocol (Weiss et al., 2003) and the Reformed Teaching Observation Protocol (RTOP; Sawada et al., 2002). These protocols consist of statements that observers typically score on a Likert scale from “not at all” to “to a great extent” and contain such statements as: “The teacher had a solid grasp of the subject matter content inherent in the lesson” (from RTOP; Sawada et al., 2002).

The RTOP in particular has been used to observe university STEM instruction. For example, it has been used to evaluate university-level courses at several different institutions to measure the effectiveness of faculty professional development workshops (Ebert-May et al., 2011) and to compare physics instructors in a study examining coteaching as a method to help new faculty develop learner-centered teaching practices (Henderson et al., 2011). The RTOP is also being used to characterize classroom practices in many institutions and in all levels of geoscience classes (Classroom Observation Project, 2011).

The RTOP was found to be unsuitable for the UBC and UMaine programs for two main reasons. The first is that the protocol involves many observational judgments that can be awkward to share with the instructor and/or the larger university community. The second is that observers must complete a multiday training program to achieve acceptable interrater reliability (IRR; Sawada et al., 2002).

More recently, new observation protocols have been developed that describe instructional practices without any judgment as to whether or not the practices are effective or aligned with specific pedagogic strategies. These observation protocols use a series of codes to characterize instructor and/or student behaviors in the classroom; observers indicate how often each behavior occurs during a class period (Hora et al., 2013; West et al., 2013). One observation protocol in particular, the Teaching Dimensions Observation Protocol (TDOP), was expressly developed to observe postsecondary nonlaboratory courses. For this protocol, observers document classroom behaviors in 2-min intervals throughout the duration of the class session (Hora et al., 2013). The possible classroom behaviors are described in 46 codes in six categories, and observers make a checkmark when any of the behaviors occur.

The TDOP instrument avoids the judgment issues associated with the RTOP, but it still requires substantial training, as one might expect for a protocol that was designed to be a complex research instrument. Preliminary work suggests that, after a 3-day training session, observers have acceptable IRR scores when using the TDOP (Hora et al., 2013). Observers at our institutions tried using this instrument, but without the full training, they found it difficult to use the TDOP in a reliable way, due to the complexity of the items being coded and the large number of possible behavior codes. We also found that the particular research questions it was designed to address did not entirely align with our needs. For example, it covers some aspects that are not necessary for faculty observation programs, such as whether an instructor uses instructional artifacts (e.g., a laser pointer or computer; Hora et al., 2013) and fails to capture others that are needed, such as whether an instructor encourages peer discussion along with clicker questions (Mazur, 1997; Smith et al., 2009, 2011). We also wanted to better characterize the student behaviors during the class period than the TDOP easily allowed.

Out of necessity, we created a new protocol called the Classroom Observation Protocol for Undergraduate STEM, or COPUS. Like the TDOP, this new protocol documents classroom behaviors in 2-min intervals throughout the duration of the class session, does not require observers to make judgments of teaching quality, and produces clear graphical results. However, COPUS is different in that it is limited to 25 codes in only two categories (“What the students are doing” and “What the instructor is doing”) and can be reliably used by university faculty with only 1.5 hours of training (Figure 1 has a description of the codes; the Supplemental Material includes the full protocol and coding sheet). Observers who range from STEM faculty members without a background in science education research to K–12 STEM teachers have reliably used this protocol to document instruction in undergraduate science, math, and engineering classrooms. Taken together, their results show the broad usability of COPUS.

DEVELOPMENT

The development of COPUS was an evolutionary process extending across more than 2 years, involving many iterations and extensive testing. It began at UBC, where science education specialists (SESs) who were working with science faculty on improving teaching (Wieman et al., 2010) wanted to characterize what both the students and instructors were doing during class. The SESs began testing various existing protocols, including the TDOP, in different classes at UBC in late 2011 and early 2012. The original TDOP did not meet our needs (as described above), so we iteratively modified the protocol through nine different versions. These changes resulted in a format, procedure, data structure, and coding strategy that was easy to implement on paper or electronically and convenient for analysis and display. The overall format of the observation protocol remained largely stable, but the categories and codes continued to evolve.

During the Fall term of 2012, 16 SESs, who are highly trained and experienced classroom observers, used this evolving protocol to observe a variety of courses in singles,
Figure 1. Descriptions of the COPUS student and instructor codes.

1. **Students are Doing**
   - **L** Listening to instructor/taking notes, etc.
   - **Ind** Individual thinking/problem solving. Only mark when an instructor explicitly asks students to think about a clicker question or another question/problem on their own.
   - **CG** Discuss clicker question in groups of 2 or more students
   - **WG** Working in groups on worksheet activity
   - **OG** Other assigned group activity, such as responding to instructor question
   - **AnQ** Student answering a question posed by the instructor with rest of class listening
   - **SQ** Student asks question
   - **WC** Engaged in whole class discussion by offering explanations, opinion, judgment, etc. to whole class, often facilitated by instructor
   - **Prd** Making a prediction about the outcome of demo or experiment
   - **SP** Presentation by student(s)
   - **TQ** Test or quiz
   - **W** Waiting (instructor late, working on fixing AV problems, instructor otherwise occupied, etc.)
   - **O** Other – explain in comments

2. **Instructor is Doing**
   - **Lec** Lecturing (presenting content, deriving mathematical results, presenting a problem solution, etc.)
   - **RtW** Real-time writing on board, doc. projector, etc. (often checked off along with Lec)
   - **FuP** Follow-up feedback on clicker question or activity to entire class
   - **PQ** Posing non-clicker question to students (non-rhetorical)
   - **CQ** Asking a clicker question (mark the entire time the instructor is using a clicker question, not just when first asked)
   - **AnQ** Listening to and answering student questions with entire class listening
   - **MG** Moving through class guiding ongoing student work during active learning task
   - **1o1** One-on-one extended discussion with one or a few individuals, not paying attention to the rest of the class (can be along with MG or AnQ)
   - **D/V** Showing or conducting a demo, experiment, simulation, video, or animation
   - **Adm** Administration (assign homework, return tests, etc.)
   - **W** Waiting when there is an opportunity for an instructor to be interacting with or observing/listening to student or group activities and the instructor is not doing so
   - **O** Other – explain in comments

Pairs, or trios across most of the departments in the UBC Faculty of Science (including the disciplines of biology, computer science, earth sciences, mathematics, physics, and statistics). We analyzed the SES generated observation data to identify coding disagreements and met with the SESs to discuss the evolving protocol and coding. These discussions covered observed behaviors they found difficult to code and/or hard to interpret, and other important elements of instructor or student behavior they felt were not being adequately captured. The protocol evolved through five different versions during this stage of testing and feedback. The final version had substantially simplified categories and all identified problems with the wording on the codes had been eliminated. Notably, it was quite simple to reliably code classes taught with traditional lectures, as a very small number of behaviors need to be coded. Therefore, the majority of the work went into improving the protocol so it could reliably characterize classes that had substantial and varied interactions between instructor and students and multiple student activities.

One substantial change during Fall 2012 was eliminating a category for judging the cognitive level of the activities. Observers had been asked to code the level of cognitive sophistication of current classroom activities, based on Bloom's taxonomy of educational objectives (Bloom et al., 1956). After multiple unsuccessful attempts to find a simple and reliable coding scheme that could capture this aspect of the classroom activities, we dropped this category. Our decision to drop this category is supported by recent work showing that, when faculty members write and evaluate higher-order questions, they use several criteria beyond the Bloom’s level, including: question difficulty, time required to answer the questions, whether students are using a new or well-practiced approach, and whether the questions have multiple reasonable solutions (Lemons and Lemons, 2012).

The second substantial change during this time was changing another category—coding the level of student engagement—from required to optional. Having a measure of student engagement is useful for providing feedback to the instructor and for judging the overall effectiveness of many instructional activities. With the coding of the levels of engagement simplified to only discriminating between low (0–20% of the students engaged), medium, or high (≥80% of the student engaged), some observers, particularly those who had some experience with observing levels of student engagement, could easily code engagement along with the other two categories, and there was reasonable consistency between observers. However, less-experienced observers found it quite hard to simultaneously code what the students were doing, what the instructor was doing, and the student engagement level. Also, there were difficulties with obtaining consistent coding of student engagement across all observers; the judgments were often dependent on the levels of engagement common to the specific disciplines and courses with which the observers were familiar. For this reason, the student engagement category was made optional. We recommend observers do not try to code it until after they have become experienced at coding the “What the students are doing” and “What the instructor is doing” categories.
Another recurring theme of the discussions with the SESs was the extent to which classroom observations could accurately capture the quality of instruction or the efficacy of student work. In the end, after SESs observed different classes across many disciplines, there was a consensus that accurately evaluating the quality of instruction and the efficacy of student work was generally not possible. These highly trained and experienced observers concluded that these evaluations require a high degree of training of the observer in the material and the pedagogic strategies, as well as familiarity with the student population (prior knowledge, typical classroom behaviors, etc.). We concluded that quality judgments of this type were not realistic goals for limited classroom observations carried out by STEM faculty members. Thus, the present version of COPUS captures the actions of both instructors and students, but does not attempt to judge the quality of those actions for enhancing learning.

After the completion of this development work at UBC, the COPUS was further tested by 16 K–12 teachers participating in a teacher professional development program at UMaine. The teachers used the COPUS to observe 16 undergraduate STEM courses in five different departments (biology, engineering, math, chemistry, and physics). While the teachers easily interpreted many of the codes, they found a few to be difficult and suggested additional changes. For example, the student code “Listening: paying attention/taking notes, etc.” was changed to “Listening to instructor/taking notes, etc.” The code was clarified, so observers knew they should select this code only when the students were listening to their instructor, not when students were listening to their peers. Also, new codes were added to capture behaviors the teachers thought were missing, such as the instructor code “AnQ: Listening to and answering student questions with entire class listening.”

The coding patterns of the two teacher observers in the same classroom were also compared to determine which specific codes were difficult to use consistently. An example comparing two teachers employing the student code “Ind” is shown in Figure 2. Figure 2A compares how two observers marked this code in the first iteration of testing, when it was described “Ind: Individual thinking/problem solving in response to assigned task.” Observer 2 marked this code throughout most of the class, and observer 1 marked this code intermittently. Follow-up conversations with observer 2 and other teachers indicated that some observers were marking this code throughout the duration of the class, because they assumed individual students were thinking while they were taking notes, working on questions, and so on, but other observers were not. Therefore, we clarified the code to be: “Ind: Individual thinking/problem solving. Only mark when an instructor explicitly asks students to think about a clicker question or another question/problem on their own.”

![Figure 2](attachment:image.png)
Table 1. Information on the courses observed using the final version of the COPUS

<table>
<thead>
<tr>
<th>Institution</th>
<th>Number of classes observed</th>
<th>Number of different STEM departments</th>
<th>Percentage of courses at the introductory levelb</th>
<th>Percentage of classes with &gt;100 students</th>
</tr>
</thead>
<tbody>
<tr>
<td>UBC</td>
<td>8</td>
<td>4b</td>
<td>100</td>
<td>63</td>
</tr>
<tr>
<td>UMaine</td>
<td>23</td>
<td>7c</td>
<td>96</td>
<td>35</td>
</tr>
</tbody>
</table>

*aSTEM courses at the first- and second-year levels.

*bBiology, chemistry, math, and physics.

*cBiology, molecular biology, engineering, chemistry, math, physics, and geology.

Figure 2B shows a comparison of the same observer pair, with the revised “Ind” code showing how the paired codes were now closely aligned.

In addition, the teacher observation data revealed a more general problem: there was a lower degree of consistency in coding student behaviors than in coding instructor behaviors, and the teachers used a very limited set of codes for the student behaviors. The earlier coding by the SESs had shown similar, but less dramatic, trends. We realized that this problem was due to a natural tendency of observers to focus on the instructor, combined with the fact the instructor-related codes came first on the survey form. Therefore, the protocol was changed, with the student codes viewed first, and we emphasized coding student behaviors during subsequent training sessions (see further details below in the Training section). As shown below, these changes appear to have fixed this problem.

These further revisions culminated in a final version of the COPUS. This version was tested by having the same 16 K–12 teachers use it to observe 23 UMaine STEM classes, and by having seven STEM faculty observers use it to observe eight UBC classrooms in pairs after 1.5 hours of training. Information about the types of classes observed is in Table 1. The seven UBC STEM faculty member volunteers who used the final protocol had not previously used the protocol and were not involved in the development process. Thus, the IRR of the protocol has been tested with a sample of observers with a wide range of backgrounds and perspectives. As discussed in Validity and Reliability, the IRR was high.

TRAINING

A critical design feature of the COPUS is that college and university faculty who have little or no observation protocol experience and minimal time for training can use it reliably. We summarize the training steps in the following paragraphs, and we have also included a step-by-step facilitator guide in the Supplemental Material.

The first step in the training process is to have the observers become familiar with the codes. At UBC, facilitators displayed the student and instructor codes (Figure 1) and discussed with the observers what each behavior typically looks like in the classroom. At UMaine, the teacher observers played charades. Each teacher randomly selected a code description from a hat and silently acted out the behavior. The remaining observers had the code descriptions in front of them and guessed the code. The remainder of the training was the same for both groups, with a total training duration of 2 hours for the K–12 teachers and 1.5 hours for the UBC faculty members.

Second, observers were given paper versions of the coding sheet and practiced coding a 2-min segment of a classroom video. An excerpt from the coding sheet is shown in Figure 3, and the complete coding sheet is included in the Supplemental Material. Observers often mark more than one code within a single 2-min interval. The first video we used showed an instructor making administrative announcements and lecturing while the class listened. After 2 min, the video was paused, and the group discussed which codes they selected. Because faculty at other institutions may have difficulty capturing videos for training, we have included web URLs to various video resources that can be used for training (Table 2).

The observers were then asked to form pairs and code 8 min of a video from a large-enrollment, lecture-style science class at UMaine that primarily shows an instructor lecturing and students listening, with a few questions asked by both the instructor and students. To keep the observers synchronized and ensure they were filling out a new row in the observation protocol at identical 2-min intervals, they used either cell phones set to count time up or a sand timer. At
the end of 8 min, the observers compared their codes with their partners. Next, as a large group, observers took turns stating what they coded for the students and the instructor every 2 min for the 8-min video clip. At this point, the observers talked about the relationship between a subset of the student and instructor codes. For example, if the observers check the student code “CG: Discuss clicker question,” they will also likely check the instructor code “CQ: Asking a clicker question.”

To provide the observers with practice coding a segment that has more complicated student and instructor codes, they next coded a different classroom video segment from the same large-enrollment, lecture-style science class at UMaine, but this time the camera was focused on the students. This video segment included students asking the instructor questions, students answering questions from the instructor, and clicker questions with both individual thought and peer discussion. The observers coded 2 min and then paused to discuss the codes. Then observers in pairs coded for an additional 6 min, again taking care to use synchronized 2-min increments. The observer pairs first compared their codes with their partners, and then the whole group discussed the student and instructor codes for each of the 2-min segments of the 6-min clip. At this point, the training was complete.

VALIDITY AND RELIABILITY

COPUS is intended to describe the instructor and student actions in the classroom, but it is not intended to be linked to any external criteria. Hence, the primary criterion for validity is that experts and observers with the intended background (STEM faculty and teachers) see it as describing the full range of normal classroom activities of students and instructors. That validity was established during the development process by the feedback from the SESs, the K–12 teachers, and those authors (M.S., F.J., C.W.) who have extensive experience with STEM instruction and classroom observations.

A major concern has been to ensure that there is a high level of IRR when COPUS is used after the brief period of training described above. To assess the IRR, we examined the agreement between pairs of observers as they used the final version of COPUS in STEM classes at both UBC and UMaine. The two observers sat next to each other in the classroom, so they could keep identical 2-min time increments, but the observers were instructed not to compare codes with each other.

To summarize how similarly observer pairs used each code on the final version of the COPUS, we calculated Jaccard similarity scores (Jaccard, 1901) for each code and then averaged the scores for both the UBC and UMaine observers (Table 3).

<table>
<thead>
<tr>
<th>Video resources that may be helpful for COPUS training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description of video</td>
</tr>
<tr>
<td>Demonstration, clicker questions, and lecture</td>
</tr>
<tr>
<td>Group activities and lecture</td>
</tr>
<tr>
<td>Clicker, real-time writing, and lecture</td>
</tr>
<tr>
<td>Real-time writing, asking/answering questions, and lecture</td>
</tr>
</tbody>
</table>

Table 3. Average Jaccard similarity scores for COPUS codes across all pairs observing in all courses for both UBC faculty observers and Maine K–12 teacher observers; numbers closer to 1 indicate the greatest similarity between two observers

<table>
<thead>
<tr>
<th>Student code</th>
<th>UBC</th>
<th>UMaine</th>
<th>Instructor code</th>
<th>UBC</th>
<th>UMaine</th>
</tr>
</thead>
<tbody>
<tr>
<td>L: Listening</td>
<td>0.95</td>
<td>0.96</td>
<td>Lec: Lecturing</td>
<td>0.91</td>
<td>0.92</td>
</tr>
<tr>
<td>Ind: Individual thinking/problem solving</td>
<td>0.97</td>
<td>0.91</td>
<td>RTW: Real-time writing</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td>CG: Discuss clicker question</td>
<td>0.98</td>
<td>0.97</td>
<td>FUp: Follow-up on clicker questions or activity</td>
<td>0.92</td>
<td>0.85</td>
</tr>
<tr>
<td>WG: Working in groups on worksheet activity</td>
<td>0.98</td>
<td>0.99</td>
<td>PQ: Posing nonclicker questions</td>
<td>0.86</td>
<td>0.80</td>
</tr>
<tr>
<td>OG: Other group activity</td>
<td>Not used</td>
<td>0.97</td>
<td>CQ: Asking a clicker question</td>
<td>0.93</td>
<td>0.97</td>
</tr>
<tr>
<td>AnQ: Students answer question posed by instructor</td>
<td>0.91</td>
<td>0.84</td>
<td>AnQ: Answering student questions</td>
<td>0.94</td>
<td>0.89</td>
</tr>
<tr>
<td>SQ: Student asks question</td>
<td>0.96</td>
<td>0.93</td>
<td>MG: Moving through the class</td>
<td>0.96</td>
<td>0.97</td>
</tr>
<tr>
<td>WC: Engaged in whole-class discussion</td>
<td>0.96</td>
<td>0.98</td>
<td>1o1: One-on-one discussions with students</td>
<td>0.94</td>
<td>0.96</td>
</tr>
<tr>
<td>Prd: Making a prediction about the outcome of demo or experiment</td>
<td>Not used</td>
<td>1.00</td>
<td>D/V: Conducting a demo, experiment, etc.</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>SP: Presentation by students</td>
<td>Not used</td>
<td>Not used</td>
<td>Adm: Administration</td>
<td>0.94</td>
<td>0.97</td>
</tr>
<tr>
<td>TQ: Test or quiz</td>
<td>Not used</td>
<td>Not used</td>
<td>W: Waiting</td>
<td>0.95</td>
<td>0.98</td>
</tr>
<tr>
<td>W: Waiting</td>
<td>0.99</td>
<td>0.98</td>
<td>O: Other</td>
<td>0.97</td>
<td>1.00</td>
</tr>
</tbody>
</table>

*SP: Presentation by students and “TQ: Test/quiz” were not selected in any of the observations at UBC or UMaine. This result likely occurred because when we asked UBC and UMaine faculty members if we could observe their classes, we also asked them if there was anything unusual going on in their classes that day. We avoided classes with student presentations and tests/quizzes, because these situations would limit the diversity of codes that could be selected by the observers.
For single codes, we calculated Jaccard similarity scores instead of IRR Cohen’s kappa values, because observer pairs occasionally marked the same code for every 2-min increment throughout the duration of the class. For example, in a class that is lecture-based, observers would likely mark the student code “L: Listening” for the entire time. In a case such as this, the observer opinion is defined as a constant rather than a variable, which interferes with the IRR calculation.

The equation for the Jaccard coefficient is \( T = n_c / (n_a + n_b - n_c) \), where \( n_c \) = the number of 2-min increments that are marked the same (either checked or not checked) for both observers, \( n_a \) = the number of 2-min increments that are marked the same for both observers plus 2-min increments observer 1 marked that observer 2 did not, \( n_b \) = number of 2-min increments that are marked the same for both observers plus 2-min increments observer 2 marked that observer 1 did not. For example, for the data in Figure 2B, the class period is 42 min in length, so there are 21 possible 2-min segments. The student code “Ind: Individual thinking” was marked 12 times by observers 1 and 2, not marked eight times by both observers, and marked by observer 2 one time when observer 1 did not. Therefore, the calculation is: \( 20/(20 + 21 - 20) = 0.95 \). Numbers closer to 1 indicate greater consistency between how the two observers coded the class.

Eighty-nine percent of the similarity scores are greater than 0.90, and the lowest is 0.80. These values indicate strong similarity between how two observers use each code. The lowest score for both the UBC and UMaine observers was for the instructor code “PQ: Posing nonclicker questions.” Comments from observers suggest that, when instructors were following up/giving feedback on clicker questions or activities, they of- ten posed questions to the students. Observers checked the instructor code “FU: Follow-up” to describe this behavior but stated they occasionally forgot to also select the instructor code “PQ.”

To compare observer reliability across all 25 codes in the COPUS protocol, we calculated Cohen’s kappa IRR scores using SPSS (IBM, Armonk, NY). To compute the kappa values for each observer pair, we added up the total number of times: 1) both observers put a check in the same box, 2) neither observer put a check in the same box, 3) observer 1 put a check in a box when observer 2 did not, and 4) observer 2 put a check in a box when observer 1 did not. For example, at UBC, when looking at all 25 codes in the COPUS, one observer pair had the following results: 1) both observers put a check in 83 of the same boxes, 2) neither observer put a check in 524 of the boxes, 3) observer 1 marked six boxes when observer 2 did not, and 4) observer 2 marked 12 boxes that observer 1 did not. Using data such as these, we computed the kappa score for each of the eight UBC and 23 UMaine pairs and report the average scores in Table 4. We also repeated this calculation using either the subset of 13 student or 12 instructor codes (Table 4).

The average kappa scores ranged from 0.79 to 0.87 (Table 4). These are considered to be very high values for kappa and thus indicate good IRR (Landis and Koch, 1977). Notably, the kappa values, as well as the Jaccard similarity scores, are comparably high for both UBC faculty and UMaine K–12 teacher observers, indicating that COPUS is reliable when used by observers with a range of backgrounds and 2 hours or fewer of training.

### ANALYZING COPUS DATA

To determine the prevalence of different codes in various classrooms, we added up how often each code was marked by both observers and then divided by the total number of codes shared by both observers. For example, if both observers marked “Instructor: Lecture” at the same 13 time intervals in a 50-min class period and agreed on marking 25 instructor codes total for the duration of the class, then 13/25, or 52% of the time, the lecture code occurred for the instructor.

We visualized the prevalence of the student and instructor codes using pie charts. Figure 4 shows observation results from two illustrative classes: one that is primarily lecture-based and one in which a combination of active-learning strategies are used. The latter class is clearly differentiated from the lecture-based class. This example illustrates how, at a glance, this visual representation of the COPUS results provides a highly informative characterization of the student and instructor activities in a class.

At a department- or institution-wide level, there are several ways to categorize the range of instructional styles. One of the simplest is to look at the prevalence of the student code “L: Listening to instructor/taking notes, etc.” across all courses observed, because this student code is the most indicative of student passive behavior in response to faculty lecturing (“Lec”) with or without real-time writing (“RtW”). Figure 5 shows that at both institutions the “L” code was marked 26–75% of the time. However, at UMaine, some of the classes have greater than 76% of the student codes devoted to listening. Faculty who teach these classes may benefit from professional development activities about how to design an effective active-learning classroom.

In addition, the data can be analyzed for a subset of faculty members who are using active-learning strategies, such as asking clicker questions. Thirty-eight percent of UBC and 43% of the UMaine classes that were observed used clickers. However, student code prevalence in these classes show that not all faculty members used clicker questions accompanied by recommended strategies, such as peer discussion (Mazur, 1997; Smith et al., 2009, 2011; Figure 6). Faculty members who are not allowing time for peer discussion may benefit from professional development on how to integrate peer discussion into clicker questions.

| Table 4. Average IRR kappa scores from the observations at UBC and UMaine |
|-----------------|-----------------|-----------------|
| Observers       | All codes (± SE)| Student codes (± SE) | Instructor codes (± SE) |
| Faculty observing UBC courses | 0.83 (0.03)    | 0.87 (0.04)      | 0.79 (0.04)    |
| Teachers observing UMaine courses | 0.84 (0.03)    | 0.87 (0.04)      | 0.82 (0.04)    |
Figure 4. A comparison of COPUS results from two courses that have different instructional approaches.

Figure 5. Prevalence of the student code “L: Listening” across several UBC and UMaeine classes.

Figure 6. Prevalence of student codes in four example courses that use clickers. In courses that use clickers with no or minimal peer discussion, the students are passively listening the majority of the time.
DISCUSSION AND SUMMARY

COPUS was developed because university observation programs needed a protocol to: 1) characterize the general state of teaching, 2) provide feedback to instructors who desired information about how they and their students were spending class time, and 3) identify faculty professional development needs. COPUS meets all of these goals by allowing observers with little observation protocol training and experience to reliably characterize what both faculty and students are doing in a classroom.

There are several uses for COPUS data. On an individual level, faculty members can receive pie charts with their code prevalence results (examples in Figure 4). These results provide a nonthreatening way to help faculty members evaluate how they are spending their time. We discovered that faculty members often did not have a good sense of how much time they spent on different activities during class, and found COPUS data helpful.

In addition, faculty members can use COPUS data in their tenure and promotion documents to supplement their normal documentation, which typically includes student evaluation information and a written description of classroom practices. Having observation data gives faculty members substantially more information to report about their use of active-learning strategies than is usually the case.

COPUS data can also be used to develop targeted professional development. For example, anonymized, aggregate COPUS data across all departments have been shared with the UMaine Center for Excellence in Teaching and Assessment, so workshops and extended mentoring opportunities can better target the needs of the faculty. One area in particular that will be addressed in an upcoming professional development workshop is using clickers in a way that promotes peer discussion. The idea for this workshop came about as a result of the COPUS evidence showing the prevalence of UMaine STEM classes that were using clickers but allowing no or minimal time for recommended student peer discussions (Figure 6).

Other planned uses for COPUS include carrying out systematic observations of all instructors in a department at UBC in order to characterize teaching practices. The information will be used with other measures to characterize current usage of research-based instructional practices across the department’s courses and curriculum.

In the end, the choice of observation protocol and strategy will depend on the needs of each unique situation. COPUS is easy to learn, characterizes nonjudgmentally what instructors and students are doing during a class, and provides data that can be useful for a wide range of applications, from improving an individual’s teaching or a course to comparing practices longitudinally or across courses, departments, and institutions.

ACKNOWLEDGMENTS

This work was supported at UBC through the Carl Wieman Science Education Initiative and by the National Science Foundation under grant #0962815. We are grateful for the assistance of all of the UBC SEs who contributed to the development of the survey; Lisa McDonnell and Bridgette Clarkston for running the UBC training session; MacKenzie Stetzer, Susan McKay, Erika Allison, Medea Steinman, and Joanna Meyer for helping to run the UMaine training session; Jeremy Smith for developing scripts to parse and analyze the data; the Maine K–12 teachers and UBC faculty volunteers who served as observers; and the faculty at UBC and UMaine who allowed their courses to be observed.

Approval to observe classrooms and instruction at UBC and publish results of that work is provided to the Carl Wieman Science Education Initiative by the University of British Columbia under the policy on institutional research. Approval to evaluate teacher observations of classrooms (exempt status, protocol no. 2013-02-06) was granted by the Institutional Review Board at the University of Maine.

REFERENCES


CBE—Life Sciences Education (LSE) is an online, quarterly journal owned and published by the American Society for Cell Biology (ASCB) in editorial partnership with the Genetics Society of America. The journal publishes original, previously unpublished, peer-reviewed articles on research and evaluation related to life sciences education, as well as articles about evidence-based biology instruction at all levels. The ASCB believes that biology learning encompasses diverse fields, including math, chemistry, physics, engineering, and computer science, as well as the interdisciplinary intersections of biology with these fields. One goal of the journal is to encourage teachers and instructors to view teaching and learning the way scientists view their research, as an intellectual undertaking that is informed by systematic collection, analysis, and interpretation of data related to student learning. Target audiences include those involved in education in K–12 schools, two-year colleges, four-year colleges, science centers and museums, and professional schools, including graduate students and postdoctoral researchers. All published articles are available freely online without subscription. In addition, published articles are indexed in PubMed and available through PubMed Central.

LSE is published online four times a year: March (Spring issue), June (Summer issue), September (Fall issue), and December (Winter issue). LSE also prints a highlights issue each year in December, featuring contributions selected from the four online issues. Submissions are accepted at any time. Articles are assigned to particular issues by the editors. To be included in an issue, manuscripts must be accepted in final form at least two months prior to the publication date.

**Determining the Suitability of a Manuscript for LSE**

**Articles.** LSE is a venue for biologists to disseminate their educational innovations to others who teach biology, as well as for dissemination of biology education research that is designed to generate more generalizable, basic knowledge about biology education. Thus, LSE publishes two types of articles: descriptions of research that breaks new ground in understanding biology teaching and learning and descriptions of the implementation and evaluation of educational innovations in the life sciences. Regardless of the nature of the work, articles should offer a logical, evidence-based chain of reasoning about the design and methods used to generate the findings and support the conclusions.

The design and interpretation of studies submitted for publication in LSE should fit the goals of the work. Articles about biology education research should describe how the study was designed and conducted to yield generalizable claims and should be applicable beyond a single course or program. Authors of this type of article are encouraged to draw from the diverse social science theories, methods, and findings to inform their work, and to clearly define terms and approaches that may be unfamiliar to a biologist audience.

Articles about educational innovations should describe the systematic collection and analysis of educational data and include rigorous reflection about the results with the aim of improving instruction. Such work can be limited to a single course or program, but the educational innovation should be sufficiently novel and the results sufficiently compelling to prompt other instructors to adopt or adapt it for use with their own students. Authors of this type of article must review relevant literature to demonstrate how a particular innovation is unique compared with previously published work.

Instructors interested in publishing their educational innovations in LSE should give careful thought to how they will assess student learning or other desired outcomes. Answering three questions can help guide the process of assessment: 1) What are your instructional or programmatic goals? 2) What should learners know or be able to do if you met your goals? 3) How can you measure or otherwise document whether learners know or are able to do what you intend? Documentation of intended outcomes can be accomplished through systematic analysis of data collected through diverse approaches, such as pretest/posttest, interviews, focus groups, surveys, or performance on coursework, including exams, papers, or lab reports. Authors should present their innovations in the same way that life scientists present their research: claims regarding efficacy must be supported by evidence. Articles that lack adequate assessment, assessment instruments, descriptions of assessment methods, or references to published assessment instruments or methods will be returned to authors without review.

All articles must include collection, analysis, and interpretation of educational data, which can be quantitative or qualitative in nature. In addition, LSE articles should: 1) address a clear educational problem or education research question, 2) demonstrate clear alignment among the problem or question being addressed, the design of the study or educational innovation, the claims being made, and the evidence used to support those claims, 3) describe how results are applicable or transferable to other settings, 4) be relevant to a defined audience of educators, and 5) make reference to related educational literature. Articles should include a formal Methods section, and any assessment tools (surveys, tests, assignments, interview or focus group questions, etc.) should be included as they were administered to participants as supplemental materials. The source of the assessment tool(s) should be described, including the rationale behind the selection or design of the tool(s). The online nature of the journal facilitates the inclusion of instructional materials such as syllabi, assignments, rubrics, laboratory protocols, or professional development guidelines. Science procedures, protocols, and results that are important for understanding how instruction was accomplished should be included as supplemental materials rather than in the body of the manuscript.

The following references may be useful for thinking about the design and conduct of biology education studies:


**Essays.** LSE publishes essays on timely and important topics related to biology teaching and learning, including assessment methods, student engagement, curriculum innovations, K–20 continuum, and other topics. Essays are framed by
personal experience and provide specific examples, but describe a problem or approach of general interest and may be synthetic across the work of many individuals. Appropriate foci for essays include reviews of current practices, policies, or research that have implications for biology teaching and learning, or personal perspectives on issues that are provocative or would otherwise be of widespread interest. The problem or approach should be presented within a scholarly context, citing references and resources that address the topic. If claims are to be made, there should be evidence from the literature or the authors’ own work. Although it is not a requirement, essays can include ideas for assessment or future research as appropriate. Manuscripts that include claims about the efficacy of an instructional approach should be submitted as articles. Descriptions of studies with preliminary or very limited data will not be considered.

LSE also publishes Research Methods essays that offer scholarly and practical advice on biology education research design and methods. Research Methods essays should be written as instructional pieces, identifying common and significant methodological issues. These essays should focus on a single topic, treating it with sufficient depth for readers to understand and be able to take action on the issue, while appealing to a broad audience of biology education scholars and education-interested biologists. Authors are encouraged to be creative in format. For example, Research Methods essays could be scholarly reviews punctuated by practical advice, or in-depth discussions of articles that illustrate exemplary methodological practice. The essays should be concise and accurate yet approachable, clearly defining technical terms and using biology-friendly analogies and examples to illustrate key points. Co-authorship by social scientists and biologists is encouraged.

Meeting Reports. The journal publishes occasional reports of meetings related to education in the life sciences. 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. A goal of LSE is to stimulate dialogue. LSE invites readers to make use of the Reader Comment function associated with each publication or submit Letters to the Editor. LSE requires that letters are directly responsive to an article published in the journal or that they bring a new or under-recognized issue to the attention of readers in a way that is informed by the inclusion of relevant data or references to the literature. Letters received through the normal online submission process will be reviewed by the Editor-in-Chief and published at her discretion. Appropriate topics for letters include commentary on education-related articles in this or other journals as well as opinions on more general topics of concern to the readership.

Features. Articles listed under the heading of “Features” are by invitation only. Authors interested in contributing to a feature should contact the editor.

Announcements. LSE accepts noncommercial announcements of meetings, workshops, and conferences and of funding opportunities and fellowships open to all.

Peer-Review Process

All submitted manuscripts and educational materials are subject to peer review. After a manuscript has been submitted to LSE, the Editor-in-Chief selects an editorial board member to guide the paper through the review process. Editorial board members select two reviewers to submit written evaluations. The board member will assess the peer reviews and determine whether the submission will be accepted as is, accepted with suggested revisions, temporarily rejected with suggestions for improvements before resubmission, or rejected with reasons explaining this decision. The corresponding author can usually expect an initial response within four weeks.

License and Publishing Agreement

Authors are required to sign a License and Publishing Agreement when a manuscript is accepted for publication. Under the terms of that agreement, authors retain copyright but grant the ASCB a perpetual license to publish the manuscript. Authors also grant to the general public the nonexclusive right to copy, distribute, or display the manuscript subject to the terms of the Creative Commons-Noncommercial-Share Alike 3.0 Unported license (http://creativecommons.org/licenses/by-nc-sa/3.0).

ASCB Policy on Research Misconduct by Authors

By submitting a paper to LSE, an author acknowledges that he or she is subject to the ASCB Policy on Research Misconduct by Authors. The policy is posted at www.ascb.org/files/research_misconduct.pdf.

Institutional Review Board Approval

Manuscripts reporting on studies involving human subjects must include explicit assurance that the research reported was approved by a local Institutional Review Board, including an Institutional Review Board number, unless the research is exempt from such review according to U.S. Department of Education guidelines. Prospective authors are advised that permission must be obtained in advance.

Guidelines for Preparing Articles, Essays, and Features

General Instructions. Submitted papers must be original (i.e., not published or submitted for publication elsewhere). Previously published material will not be considered. Publication of a short abstract does not constitute prior publication, nor does presentation of data at a scientific meeting or in a webcast of such a meeting. The posting of a manuscript on an author’s personal website or in an online institutional repository prior to submission of the manuscript to LSE is not considered prior publication.

The text should be written in clear, concise, and grammatical English. Manuscripts ordinarily begin with an overview of how the work presented is relevant to the classroom, laboratory, or curriculum and what student outcomes are expected. Whenever possible, incorporate materials by citing relevant publications, without repeating already published works. Manuscript files must be submitted in .doc, .docx, or .rtf format.

Tables and illustrations should convey information effectively and must be uploaded separately. Graphs and figures should be provided digitally as separate TIF or EPS files. Authors are encouraged to take advantage of the online nature of LSE. Video, audio, databases, images, animations, molecular structures, and other electronic resources may be linked as supplemental material for further consideration by readers.
International authors may wish to consider using an editorial service, such as ScienceDocs (www.sciencedocs.com), The Medical Editor (www.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 monitoring editors and reviewers (please include institution and email address). For article submissions, authors are encouraged to indicate whether their work is best described as research that aims to yield new insights about biology teaching and learning or as the implementation and evaluation of educational innovations in the life sciences.

**Title Page.** Page 1 should include the title of the manuscript, the type of manuscript being submitted (e.g., article, essay, feature, letter to the editor, response), the number of characters in the manuscript, a shortened running title (not to exceed 42 characters and spaces), and the names and affiliations (including department, institution, city, state, and zip code) of all authors in the order in which they should appear. List the corresponding author separately with complete postal and email address and telephone and fax numbers. Keywords should also appear on page 1. Include at least five keywords selected from the text of the article. If possible, include a keyword that indicates the target learners (primary, secondary, undergraduate, graduate, general public, etc.).

If one or more of the authors of a research paper that assesses the effectiveness of a product or curriculum was also involved in producing the product or curriculum, readers need to be fully aware of this potential conflict of interest. Therefore, any potential conflicts of interest should be clearly stated on the title page of the manuscript. The author and the product should be identified, and a statement included that no promotion of a particular product to the exclusion of other similar products should be construed. This will be noted under the byline if the manuscript is accepted for publication.

**Abstract.** Page 2 should contain the abstract, which should be no more than 200 words long and should summarize the important points in the manuscript.

**Manuscript Text.** The text of the paper should begin on page 3. LSE follows the style guidelines of the Council of Biology Editors Style Manual. For chemical nomenclature, follow the Subject Index of Chemical Abstracts. Capitalize trade names and give manufacturer names and addresses. Do not include figures or tables within the body of the manuscript. A format of Introduction, Methods, Results, Discussion, and References is encouraged, but other formats may be more appropriate for some topics. Manuscripts should include line and page numbers.

**Accessing Materials.** Describe how to access new educational materials if the study or use of such materials is the subject of the paper. If materials are online, provide a URL to the material. Any registration requirements or agreements inherent in the use of the materials should be described. If there are no online materials, simply state “No additional materials available online.” For other new educational materials presented in the manuscript, authors should describe how readers can access the materials, what format is available (e.g., DVD, CD-ROM, PDF files, and html pages), how to request copies, and if there are any costs. LSE encourages provision of materials on a nonprofit basis, but recognizes that this is not always feasible. Please contact LSE with any questions regarding this policy.

**Acknowledgments.** Identify financial sources and other sources of support for the research being reported in the manuscript.

**References.** Place the reference list immediately following the manuscript text (beginning on a new page). Consult the most recent issue of LSE for reference citation formats. Cite references in the text alphabetically by name and date (e.g., Jones and Smith, 1987; Smith et al., 1988). Only published articles or manuscripts accepted for publication can be listed in the Reference section. References should contain complete titles and inclusive page numbers and should be listed in alphabetical order. Abbreviate the names of journals as in PubMed. The abbreviations can be found at LinkOut Journal Lists. Reference citations may include up to 10 authors followed by et al.

Citations to online websites are acceptable reference links. In the text, include the name of author for the cited website and the date accessed (e.g., National Center for Biotechnology Information, 2005). The citation should include the author (if known), date of latest modification (if known) or copyright, title of page, title of complete work (if appropriate), URL (e.g., www.website.org), and date of access. For example: National Center for Biotechnology Information (2005). NCBI Home Page. www.ncbi.org (accessed 5 February 2005).

Unpublished results, including personal communications and submitted manuscripts, should be cited as such in the text. Personal communications must be accompanied by permission letters unless they are from the authors’ own work.

**Footnotes.** Call out footnotes at the appropriate place in the text with a superscript numeral. The footnote text should be placed on a separate page after the References.

**Figures.** All figures should be uploaded as individual files.
**Figure Legends.** Figure legends should appear in numerical order after the References. Figure legends should provide an overview of the figure and details that describe any component parts.

**Tables.** All tables must be cited in order in the text of the manuscript. Individual table files need to be uploaded separately.

**Supplemental Material.** Upload all supplemental material (except in the case of videos) together in one combined PDF. Be sure to upload the final version of the supplemental material. It will be posted online as received and will not be edited.

**Guidelines for Preparing Book Reviews**

Book reviews should include the complete book title, publisher information, author detail, ISBN, and retail price. The reviewer should suggest a title for the book review, but the title should be informative and descriptive of the book to most readers.

**Guidelines for Preparing Digital Artwork**

Digital artwork must accompany the manuscript submission. Figures should be uploaded as separate files with the manuscript through the online manuscript submission system. Because artwork must be of sufficient quality for print reproduction, *LSE* asks that all artwork be prepared using professional graphic art software. Word processing and presentation software packages (such as Word and PowerPoint) are inadequate for preparing high-quality digital artwork.

**Prepare all digital artwork as RGB TIF images, at 300 dpi resolution, or EPS images:**

**Figure Size.** Prepare figures at the size they are to be published.

- Up to 1 column wide: Figure width should be 4.23–8.47 cm.
- 1 to 1.5 columns wide: Figure width should be 10.16–11.43 cm.
- 2 columns wide: Figure width should be 14.39–17.57 cm.

The figure depth must be less than or equal to 23.5 cm.

**Guidelines for Preparing Electronic Resources**

It is possible for authors to submit for peer review electronic works including, but not limited to, animations, Chime tutorials, movies, interactive websites that may include quizzes, images (electron micrographs, photomicrographs, etc.), Java Applets, searchable databases, etc. Articles that describe new educational uses of existing resources (e.g., Expression Connection, FlyBase, Database of Interacting Proteins, etc.) are also of interest. Manuscripts of this type should provide detailed instructions for use of the resource by the target audience.

For works using someone else’s electronic resource (such as a database), a letter from the creator or curator of the resource indicating their willingness to support free pedagogical use of their work must be included. The same general rules for evaluation will apply to all electronic submissions. All submissions will be evaluated for (1) pedagogical content, (2) clear description of goals and expected student outcomes, (3) transferability to other settings, (4) appropriateness for the target audience, and (5) references to related educational literature. After publication of the electronic work(s), the authors will be encouraged to submit the work to other databases (National Digital Library, BEN, etc.); however, it is expected that the *LSE* publication citation will remain associated with the work. This will allow viewers to read a more in-depth discussion of the work.

All such electronic works must be freely available, and will be hosted on the *LSE* server or on the ASCB server, with the exception of large databases. This will ensure stable access to the works with a nonchanging URL.

A manuscript should accompany any electronic submission. The manuscript should describe (1) the learning goals or purpose of the electronic work, (2) the target audience, (3) development of the electronic work (describe hardware and software used), (4) platform availability (see below), (5) a description of any necessary hardware or software, with links to the appropriate sites for downloading (e.g., plugins, helper applications, etc.), and (6) assessment of the work’s impact on student learning.

Ideally, submitted works should work on any platform (PC, Mac, Unix) and on all browsers. If there are known restrictions, these should be included in the manuscript. *LSE* can help authors test their works for such limitations if they do not have access to certain platforms or browsers.

The electronic work may have been hosted previously on any website, but the authors may not have previously published any description of the electronic work other than the associated Web pages. Published journal descriptions of the electronic work will preclude publication in *LSE*, with the exception of abstracts or presentations at professional meetings.

The online publication will include hyperlinks to the work that will appear in a new browser window, if appropriate. This capacity could be helpful to the authors since they could provide directions for readers as needed to illustrate particular aspects of the work. The layout and submission process for the manuscript accompanying an electronic work should follow the same general format as other categories. The electronic work should be submitted to *LSE* at the same time as the manuscript. If this presents a problem, contact *LSE* staff for assistance.

**Sharing Materials and Data**

Publication of a manuscript in *LSE* implies that the authors agree to make available, to the extent legally permissible, all propagative materials such as mutant organisms, cell lines, recombinant plasmids, vectors, viruses, monoclonal antibodies, instructional materials, and assessment instruments that were used to obtain results presented in the article. Prior to obtaining these materials, interested scientists will provide the authors with a written statement that they will be used for noncommercial research purposes only. Authors are encouraged to share raw data with qualified researchers who wish to reproduce or further analyze the authors’ work for noncommercial purposes. Sharing of data on human subjects should be consistent with the conditions of the Human Subjects protocol approved by the authors’ Institutional Review Board (and any other agreements made with the subjects) for the work reported.

**How to Submit Manuscripts**

**Electronic Submission.** Authors must submit manuscripts online at www.cellbiologyeducation.org. Specific instructions...
on how to submit your manuscript are available at the submission site. Authors should submit the manuscript in .doc, .docx, or .rtf as two text files, the first containing the cover letter and the second containing the manuscript file. Figures, tables, and Supplemental Material, including Appendixes, must be submitted as individual files. Do not embed figures in the manuscript. If you are submitting a feature that does not contain an abstract or keywords, write “There is no abstract” in the required abstract field and “one,” “two,” and “three” in the keyword fields.

Questions regarding submission guidelines can be directed to: cbe@ascb.org or 301-347-9304.

General Questions

At any stage in the submission process, authors with questions should contact the LSE Editorial Office at 301-347-9304 (phone); 301-347-9350 (fax); cbe@ascb.org; or The American Society for Cell Biology, 8120 Woodmont Avenue, Suite 750, Bethesda, MD 20814–2762.

You are encouraged to contact editorial board members or the Editor-in-Chief by email to discuss submissions to LSE.

The LSE website is an additional resource to authors. See earlier issues of LSE for examples of the different types of manuscripts published.
Learn science from leading researchers

Go behind the scenes of the world of biology

Hear scientists tell stories of discovery, give career advice, share ideas, and more

Best of all, it’s free! Visit iBiology.org

Supported by:
We believe in the power of individuals to advance science through research and science education, making discoveries that benefit humanity.

UNDERGRADUATE PROGRAMS

**HHMI initiatives encourage faculty to teach science more effectively and enable undergraduate students from diverse populations to experience authentic research early in their academic careers**

**Colleges and Universities** – institutional grants to advance undergraduate science education

**Science Education Alliance** – a nationwide research-based course for beginning undergraduate students

**HHMI Professors** – grants to research university faculty members who apply the creativity and scholarship that made them successful in research to important challenges in science education

**Exceptional Research Opportunities Program (EXROP)** – places highly motivated undergraduates from groups underrepresented in the sciences into intensive 10-week summer research experiences with HHMI scientists

**National Math and Science Initiative** – a grant to support research universities in the UTeach program, which prepares science and math undergraduates for careers as K-12 teachers
GRADUATE PROGRAMS

*HHMI initiatives create opportunities for students from diverse populations to develop into the leaders of science, including faculty members whose responsibility is to prepare the next generation of scientists and educators*

*Gilliam Fellowships for Advanced Study* – support for PhD students who will contribute to the diversity of scientific leadership

*International Graduate Student Research Fellowships* – support for PhD students from abroad who are studying in the U.S.

*Advanced Courses* – support for advanced training courses at Cold Spring Harbor Laboratory, Jackson Laboratory, and the Marine Biological Laboratory

PHYSICIAN SCIENTIST TRAINING

*HHMI supports medical, veterinary, and dental students interested in careers as physician scientists*

*Medical Research Fellows* – a year-long full-time research experience for medical, veterinary, and dental students

*Summer Medical Fellows* – support for medical, veterinary, and dental students who engage in a 10-week intensive research experience with HHMI scientists

www.hhmi.org/ugp
ARTICLES

Is Peer Interaction Necessary for Optimal Active Learning?
Debra L. Linton, Jan Keith Farmer, and Ernie Peterson ................................................................. 55–64

Integrating Quantitative Thinking into an Introductory Biology Course Improves Students’
Mathematical Reasoning in Biological Contexts
Susan Hester, Sanlyn Buxner, Lisa Elfring, and Lisa Nagy ............................................................... 65–75

Identification of Threshold Concepts for Biochemistry
Jennifer Loertscher, David Green, Jennifer E. Lewis, Sara Lin, and Vicky Minderhout .................. 76–88

Redefining Authentic Research Experiences in Introductory Biology Laboratories and Barriers
to Their Implementation
Rachelle M. Spell, Judith A. Guinan, Kristen R. Miller, and Christopher W. Beck ......................... 89–97

Getting Under the Hood: How and for Whom Does Increasing Course Structure Work?
Sarah L. Eddy and Kelly A. Hogan ......................................................................................................... 98–113

What Do I Want to Be with My PhD? The Roles of Personal Values and Structural Dynamics
in Shaping the Career Interests of Recent Biomedical Science PhD Graduates
Kenneth D. Gibbs, Jr., and Kimberly A. Griffin .................................................................................. 114–126

Development of the Biology Card Sorting Task to Measure Conceptual Expertise in Biology
Julia I. Smith, Elijah D. Combs, Paul H. Nagami, Valerie M. Alto, Henry G. Goh,
Muryam A. A. Gourdet, Christina M. Hough, Ashley E. Nickell, Adrian G. Peer,
John D. Coley, and Kimberly D. Tanner ............................................................................................... 127–143

Understanding Clicker Discussions: Student Reasoning and the Impact of Instructional Cues
Jennifer K. Knight, Sarah B. Wise, and Katelyn M. Southard .............................................................. 144–153

The Classroom Observation Protocol for Undergraduate STEM (COPUS): A New
Instrument to Characterize University STEM Classroom Practices
Michelle K. Smith, Francis H. M. Jones, Sarah L. Gilbert, and Carl E. Wieman ................................. 154–163

The journal is supported in part by an Undergraduate Science Education Program grant from the Howard Hughes Medical Institute.

8120 Woodmont Ave., Suite 750, Bethesda, MD 20814, USA; Phone: 301-347-9300; Fax: 301-347-9310; www.ascb.org