From CREATE Workshop to Course Implementation: Examining Downstream Impacts on Teaching Practices and Student Learning at 4-Year Institutions

KRISTY L. KENYON, BRADLEY J. COSENTINO, ALAN J. GOTTESMAN, MORGAN E. ONORATO, JAMILA HOQUE, AND SALLY G. HOSKINS

The faculty workshop model has long been used for disseminating innovative methods in STEM education. Despite significant investments by researchers and funding agencies, there is a dearth of evidence regarding downstream impacts of faculty development. CREATE is an evidence-based strategy for teaching science using primary literature. In this study, we examined whether workshop-trained faculty applied CREATE methods effectively and whether their students achieved either cognitive or affective gains. We followed 10 workshop alumni at different 4-year institutions throughout the United States. External observations of the teaching indicated a high fidelity of CREATE implementation. The students made significant gains in cognitive (e.g., designing experiments) and affective (e.g., self-efficacy in science process skills) domains. Some student outcomes correlated with particular characteristics (e.g., class size) but not with others (e.g., teaching experience). These findings provide evidence for the robustness of the CREATE dissemination model and provide perspective on factors that may influence pedagogical reform efforts.

Keywords: CREATE, primary literature, teaching, learning, STEM

National movements have made some headway toward inspiring curricular changes in science education (AAAS 2015). Although more voices may be championing the call for educational reforms, much remains unknown about the depth and breadth of pedagogical shifts in response to such efforts (Henderson et al. 2011, Smith et al. 2014, Wieman and Gilbert 2014). Traditional, teacher-focused instruction is still the norm for a majority of undergraduate STEM courses (Yager 2013, Walter et al. 2016).

Of central importance for reform efforts is to understand how faculty learn to teach and what factors influence their practices. Explicit pedagogical training is highly variable across science graduate programs (Tanner and Allen 2006, Love Stowell et al. 2015), even though academic positions require teaching. Decisions about teaching are often influenced by personal attitudes rather than pedagogical evidence (Andrews and Lemons 2015, Turpen et al. 2016). Competing priorities (e.g., establishing a funded research program) may contribute to faculty de-emphasizing teaching, especially if institutional policies reify ideas about its value (or lack thereof).

Workshops and formal mentoring programs are the primary means for sharing evidence-based teaching methods (Derting and Ebert-May 2010, 2011, Henderson et al. 2012, Stevens and Hoskins 2014, 2015). Faculty development workshops have produced positive shifts in participants' views and intentions on teaching with innovative strategies (e.g., First IV, Ebert-May et al. 2015; CREATE, Stevens and Hoskins 2014). In a study of physics faculty, Henderson and colleagues (2012) determined that professional workshops are effective in the early phases of the innovation-decision model (Rogers 1995): Participants can learn new teaching methods and be convinced of potential benefits. After training, the implementation and continuation phases present greater challenges (Dancy et al. 2016, Stains and Vickery 2017). Workshop-trained faculty may believe themselves

to be applying lessons learned when, in fact, they are not (Ebert-May et al. 2011, Dancy et al. 2016). Methods that are effective when applied by education specialists can be less so when used by motivated but inexperienced faculty (Andrews et al. 2011). Some faculty need several iterations of teaching before they are able to apply workshop training effectively (Pfund et al. 2009). As was noted by Derting and colleagues (2016), "Studies that generate reliable evidence of transfer of training following professional development are needed. Research on transfer—that is, do people apply to their job what they learned in training? (Yelon et al. 2014)—merits empirical evidence of whether and how professional development affects subsequent teaching practices."

To fully understand the downstream effects of professional development requires an examination of teaching and learning outcomes associated with the implementation of training. Factors such as status, teaching experience, academic setting, the type of course, student resistance, time, and personal preferences may affect faculty decisions about applying new teaching methods (Coil et al. 2010, Andrews and Lemons 2015). Workshop-trained faculty may reinvent learned strategies that might influence the efficacy of a specific teaching method (Henderson et al. 2011). Ultimately, the success of faculty professional development is contingent on its impact on student learning (Stevens and Hoskins 2014, Van Vliet et al. 2015, Connell et al. 2016, Kenyon et al. 2016).

Examining the downstream effects of CREATE workshop training

CREATE (for *consider, read, elucidate hypotheses, analyze, and interpret data; think of the next experiment*) is a methodology that situates course content and learning within the context of the primary scientific literature (Hoskins et al. 2007). In CREATE pedagogy, the basic curricular unit is a module of linked research articles (e.g., a series of articles produced by a single laboratory). Students apply different pedagogical tools in deconstructing each article in a systematic way (supplemental table S1). In the first step, the students read introductory paragraphs, identify key ideas, and then produce a concept map reflecting the central question or hypothesis of the study. In generating a concept map, the students must harness previous knowledge of scientific concepts in order to construct new knowledge (Novak 2003). Second, the students sketch drawings to illustrate observational or experimental methods. Drawing visual representations requires a range of cognitive processes and facilitates model-based reasoning (Quillin and Thomas 2015). Next, the students critically analyze the data shown in figures and tables using annotation, transformation (e.g., converting data in table form into a figure), and text translation. Assignments based on the CREATE toolkit prepare students for in-class activities involving small group work (e.g., generating a consensus map), data presentations, and timely mini-lectures on specific concepts. Following deep discussions of the article, the students are asked to propose next experiments that would extend the research in a new direction. In a final step, the students participate in a grant panel activity modeled on the peer review system of funding agencies. The students practice skills of communication and argumentation as they debate, critique, and rank order the written proposals. This activity also provides a platform for discussions on the nature of science, ethics, and public policies governing science. Once the students complete the first cycle, they build on knowledge and skills through a reiterative process as they analyze related articles within the module. Therefore, in CREATE-based courses, students learn how to consolidate and apply content knowledge as they develop the skills of analysis, experimental design, and argumentation (Hoskins and Krufka 2015).

The primary site of CREATE development and implementation was a minority-serving institution (Hoskins et al. 2007, 2011, Gottesman and Hoskins 2013). Subsequently, we designed faculty development workshops for dissemination. During these workshops, faculty participants experience CREATE-based instruction firsthand. The participants learn foundational principles through activities that model the different steps of the CREATE strategy. We challenge the faculty to complete tasks such concept mapping, drawing sketches of experimental methods (or observational study), and designing next-step experiments. The workshop sessions reflect a typical CREATE learning environment, with an emphasis on collaborative interactions and discussion. During the final phase, we guide the participants as they develop CREATE teaching materials for their own courses. The trainees also practice teaching with CREATE, whereby one participant leads a mock class session with other workshop participants serving as the students. We provide critical feedback, and the "student" participants offer ideas for modifications. CREATE workshops have been shown to positively affect faculty views on teaching, learning, and intended practices (Hoskins et al. 2017).

As was noted above, a central question facing science educators is whether faculty development translates into meaningful changes in teaching and student learning outcomes. To date, we have investigated a limited group of secondary CREATE implementations at 4-year and 2-year institutions following workshop training (Stevens and Hoskins 2014, Kenyon et al. 2016). Our findings indicate that CREATE courses promote cognitive and affective gains for students in a variety of academic settings. In this new study, we probed more deeply into the postworkshop outcomes of new CREATE implementations at 10 4-year institutions throughout the United States. We focused on the following questions: Do faculty implementers apply CREATE pedagogy in alignment with workshop training? Do students demonstrate cognitive and affective gains in CREATE courses, consistent with previous findings? Are student cognitive outcomes in CREATE courses related to faculty characteristics (e.g., tenure status, teaching experience) or situational

Table 1. Institutional and student demographics of 4-year implementations.

Institution[a]	Carnegie classification	Carnegie undergraduate profile	Course (level)	Number of students			Number of students in each major
				Total	Female[b]	Male[b]	
1	Doctoral universities: higher research activity	4-year, medium full-time, *selective*, higher transfer in	Biology (immunology, upper level)	12	7	5	Biology, 11; Other, 1
2	Doctoral universities: higher research activity	4-year, full-time, *more selective*, higher transfer in	Biology (microbial ecology, intermediate, upper level)	19	8	11	Microbiology, 7; Molecular Biology 3; Botany, 1; Chemistry, 1
3	Master's colleges and universities: larger programs	4-year, full-time, *selective*, higher transfer in	Biology (general biology, intermediate or upper level)	15	12	3	Biology, 14; Premed, 1
4	Baccalaureate colleges: diverse fields	4-year, full-time, *selective*, higher transfer in	Psychology (introductory)	26	15	11	Undecided, 5; Sciences, 7; Pyschology, 1; Education, 4; Social Sciences, 5; Business, 4;
5	Master's colleges and universities: larger programs	4-year, full-time, *selective*, higher transfer in	Biology (microbiology, intermediate or upper level)	16	8	4	Biology, 9; Forensics, 1; Animal Science, 1; Nursing, 1;
6	Baccalaureate colleges: arts and sciences focus	4-year, full-time, *more selective*, lower transfer in	Biology (conservation biology, intermediate or upper level)	27	15	12	Undeclared, 1; Biology, 22; Environmental Sciences, 4;
7	Master's colleges and universities: larger programs	4-year, full-time, *selective*, higher transfer in	Biology (virology, intermediate or upper level)	20	10	10	Biology, 17; Other Sciences, 3;
8	Doctoral universities: highest research activity	4-year, medium full-time, *selective*, lower transfer in	Biology (ecology and evoleution, intermediate or upper level)	38	22	16	Ecology or Evolutionary Biology or Biology, 35; Other, 3;
9	Doctoral universities: moderate research activity	4-year, full-time, *more selective*, lower transfer in	Biology (genetics, intermediate or upper level)	33	17	16	Biotechnology, 14; Molecular Biology, 5; Biological Medicine, 5; Other Biology, 5;
10	Baccalaureate colleges: arts and sciences focus	4-year, full-time, *more selective*, lower transfer in	Chemistry (intermediate or upper level)	15	15	0	Biochemistry, 12; Premed, 2; Agricultural Sciences, 1;

[a]Eight states are represented in this study.
[b]The self-identified gender on the OE student survey. The total number of students based on OE data was 221.

factors specific to the learning environment (e.g., class size; Carnegie selectivity rating)?

The design included an examination of teaching practices (by an independent, trained specialist) and assessments of students in both cognitive and affective realms. We addressed the reproducibility of previous findings from earlier CREATE implementations (e.g., Stevens and Hoskins 2014). We also used linear mixed models to examine whether course-specific or faculty characteristics correlate with student outcomes on cognitive assessments.

Training and selection of CREATE implementers

We conducted four CREATE in-residence workshops (4.5 days) for 2-year and 4-year faculty from institutions throughout the United States. The workshops were held at Hobart and William Smith Colleges (Geneva, New York). In total, we trained 63 faculty from 29 4-year institutions (in 30 states) and 1 Canadian. After the workshop, approximately 30% of the workshop participants submitted written applications for the implementation phase

of the study. Funding constraints limited selection to 10 4-year faculty (and an equal number of 2-year faculty; see Kenyon et al. 2016). We considered factors such as faculty status (e.g., tenure), course type, geography, and differences in student populations in choosing implementers. Seven untenured and three tenured faculty were selected (table S1): eight in biology (different areas), one each in psychology and chemistry (table 1). Years of teaching experience did not align with tenure status; half of the cohort had taught for 6 years or more (table S1). The colleges and universities varied in type and in student demographics (table 1).

The faculty implementers worked with the present article's principal investigators (PIs) to develop teaching materials and obtain local IRB approval. They also administered all student assessments, participated in periodic conference calls with the PIs, scheduled visits from the outside evaluator, and completed postobservation interviews. The implementers received a stipend for their efforts (after the course was complete).

External evaluation of implementer teaching

Marlene Hurley served as the outside evaluator (OE) for previous CREATE studies (Stevens and Hoskins 2014, Kenyon et al. 2016). She observed and evaluated all four summer workshops and the 10 implementations of this study. To this end, Hurley visited each implementer's campus twice, at different points in the semester or term. The first visit occurred early, and she observed one class session and gave student surveys during this first visit. During the second visit, she observed two class sessions, gave post-assessments, and interviewed the implementer.

For classroom observations, Hurley used a modified observation protocol developed by Iris R. Weiss (https://www.nsf.gov/pubs/1997/nsf97153/c3app_a.htm) for the National Science Foundation on the basis of research stemming from the Local Systemic Change through Teacher Enhancement Initiative program (Baniflower et al. 2007). The OE recorded time spent on teacher-focused activities (e.g., lecture) and student-focused activities (individual, pairs, group). In addition, she evaluated each class session on four distinct areas crucial to the CREATE strategy: the design of the session (18 statements), the instruction of the session (27 statements), the nature of science (13 statements, including one statement tracking the number of CREATE tools used), and the science content in the session (6 statements). The OE rated each session using a 5-point scale for each statement (1, *not at all toward*; 5, *to a great extent*; see the supplemental materials for the full protocol). These data were compiled into a synthesis rating for each focus area, using descriptive statistics. The OE calculated a mean rating per implementer (averaging scores from three sessions) to compare teaching across institutions (see the supplemental materials for the OE's methods and protocols). The OE provided all reported data in a final summary report submitted to the PIs (Marlene Hurley, personal communication, December 2014).

Student cognitive and affective assessments

We probed student outcomes in CREATE courses using three cognitive tests and one affective survey. The faculty implementers invited their students to participate on a voluntary basis in an effort to improve undergraduate education in science. The students choosing to opt out were provided course-related material to read during the assessment time. Anonymity was ensured on all assessments. The participating students were asked to use nonidentifying code numbers on their tests and surveys, allowing the PIs to match pre- and postcourse responses.

The Experimental Design and Ability Test (EDAT; Sirum and Humburg 2011) challenges students to design a way to test the credibility of a claim regarding the benefits of a nutritional supplements (ginseng). The students' open-ended written responses were analyzed by at least two individuals (blind with respect to the status—before or after treatment) using the 10-point rubric published by Sirum and Humburg (2011). The scores from a subset of 20 students were compared to determine initial concordance. The scorers reconciled minor differences and then achieved a high level of agreement in another subset of 20 students (Pearson's $r >$ .8; Best and Kahn 2006). The scores from the two individuals were subsequently averaged for each student response.

The Critical Thinking Test (CTT) is an instrument originally adapted from the *Field-Tested Learning Assessment Guide* (www.flaguide.org). Three questions are derived from the "General Science/Conceptual Diagnostic Test/ Fault Finding and Fixing/Interpreting and Misinterpreting Data" section of the FLAG site, and a fourth question was designed in a previous CREATE study (Hoskins et al. 2007) to focus on biological data analysis (see the supplemental materials for a full version). Each question requires the student to evaluate a set of data and an associated conclusion. The students were prompted to write open-ended responses explaining either their agreement or their disagreement with the stated conclusion, using data to support their view. Their responses were assessed by two independent scorers, who were blind to designation as before or after treatment. Logical and illogical statements were quantified as has been done previously (Hoskins et al. 2007, Stevens and Hoskins 2014), with a small subset initially compared and reconciled for concordance. The final scores from the two independent scorers were then averaged for each student response.

The concept map (c-map) test is a closed-book assessment specific to the course's content. Concept mapping is a cognitive task that requires a testee to organize, relate, and define relationships among discrete units of knowledge (Novak 2003). For the c-map assessment, each implementer provided five seed terms specific to their course's context. Their students were given 15 minutes to construct a closed-book concept map using the seed terms and any other terms they wished to add. The pretreatment c-map assessment was given early on, but after the students had learned and practiced concept mapping. The posttreatment c-map assessment used the same seed terms and was given near the end of the course (closed book). All of the student maps were scored for the number of concepts, links, labelled links, orphan concepts (concepts with no links), and dead-ends (concepts linked to a single other concept). After the course, we invited the implementers to assess the links on their students' maps, and four agreed to do so. We concealed the status (before or after treatment) on all maps, and each implementer quantified the number of valid or flawed links on the basis of course content.

The Survey of Student Attitudes, Abilities, and Beliefs (SAAB) survey was designed to evaluate multiple categories of self-rated abilities and confidence in skills (e.g., *I am confident in my ability to critically review scientific literature*) and epistemological beliefs (e.g., the data from a scientific experiment can only be interpreted in one way). The SAAB survey was developed previously through factor analysis of data obtained at a 4-year institution (Hoskins et al. 2011).

Table 2. OE evaluation of class sessions.

Institution	Teacher-centered focus (percentage)	Student-centered focus (percentage)
1	45.5	54.5
2	29.5	70.5
3	24.4	75.6
4	50	50
5	66.7	33.3
6	20	80
7	20	80
8	50	50
9	17	83
10	41.7	58.3

Note: The OE tracked and timed the activities and interactions for each class session. The categories for activities included *lecture*, *student presentations*, *hands on*, *minds on*, *reading*, *writing*, *assessment*. The scores were averaged over three class sessions (1–7, 9, 10), or in a single case (8), two sessions.

Statistical analysis of student data

We used paired *t*-tests to assess changes in student scores before and after implementation. Paired *t*-tests were used because the distribution of difference scores was largely normal (supplemental figure S1). Some student responses were eliminated when those students were not represented across all three assessments (EDAT, CTT, SAAB); the selected data reflected the majority of those in each course (supplemental table S2; the data are available at *DRYAD/datadryad.org*). To account for the lack of independence among the students' test scores within classes, we first averaged across difference scores (posttreatment – pretreatment score) for students within each class (see the supplemental materials for institutional variation). We then tested whether the mean difference score across classes was consistent with a null hypothesized value of 0 ($n = 9$ classes for CTT and SAAB test metrics; $n = 8$ for EDAT and c-map; the difference in sample size was due to an implementer not giving either the EDAT or the c-map assessment). The *t*-tests were conducted in R (R Core Team 2016).

We used linear mixed models to examine how changes in test scores were related to faculty characteristics, class size, and institutional setting (see the supplemental materials for the full details). Faculty status was defined as either tenured or not; experience level was classified as more or less than 6 years of teaching (table S1). We made a distinction between tenure and experience level because not all implementers were in tenure-track positions, and we used 6 years as the threshold for experience level because this aligns with the typical time before a tenure decision. Class size was determined from the OE observation data (table 1). Because the students' participation was anonymous, we did not have access to the students' grades or other personal attributes. As a general proxy for institutional differences in the student populations, we used the selectivity rating of the Carnegie Classification of Institutions of Higher Education (*http://carnegieclassifications.iu.edu*; table 1). Carnegie metrics use entrance exam scores (SAT, ACT) to define institutions as inclusive, selective, or more selective. Six of our implementations were at institutions categorized as selective, and four were at more selective institutions.

Finding 1: External evaluation indicates faculty effectively applied CREATE in course implementations. The OE determined that most of the implementers taught sessions that were predominantly student centered, with minimal time spent on teacher-focused activities such as lecturing (table 2). There was no obvious relationship between class size and teacher or student activity (table 1). As a cohort, the faculty implementers achieved high ratings (e.g., a range of 3.90–4.40 on a 5-point scale) across the four categories assessed (table 3). The implementers consistently taught with the CREATE toolkit, using a range of CREATE-related activities and techniques across three class sessions (10–17 per session, data not shown; M. Hurley). Increasing student interactions, group work, and activities encourages a deeper investment in the learning environment, arguably supporting greater gains in student performance (e.g., Freeman et al. 2014, Connell et al. 2016).

Our external evaluator used an observation protocol (by I. Weiss) grounded in the Local Systemic Change through Teacher Enhancement Initiative (Baniflower et al. 2007). Her observations captured the degree of constructivism embedded within specific applications of CREATE pedagogy, as well as details identifying the quantity or type of activities and interactions that occurred. We note that the OE protocol was consistent with other instruments used in science education studies (e.g., the Reformed Teaching Observation Protocol [RTOP], Sawada et al. 2002; Classroom Observation Protocol in Undergraduate STEM [COPUS], Smith et al. 2013). We acknowledge some limitations in the OE evaluation of teaching. The faculty implementers and the students may have altered their behaviors when the OE was present; a similar concern arises when students are videotaped for other types of teaching evaluations (e.g., RTOP; Ebert-May et al. 2011, 2015, Smith et al. 2013, Lund et al. 2015, Derting et al. 2016). Another potential concern is that there was a single OE. However, we note that there are challenges associated with multiple evaluators applying instruments such as RTOP, especially for summative assessments of teaching (e.g., Amrein-Beardsley and Osborn Popp 2012). Moreover, we assert that the OE of this study was uniquely qualified to assess all 10 implementations, because she had deep knowledge of CREATE workshop training and was not affiliated with any of the institutions. The OE traveled to each site at least twice, observing three sessions (consecutive at the second visit). Lund and colleagues (2015) have argued that it takes a week of observations to determine faculty teaching styles. The OE likely captured nuances of classroom dynamics unique to real-time assessment.

Table 3. External teaching evaluation of 4-year implementers.

Focus area	Statements of assessment	Mean	Standard deviation
Design of session	The session design considered student attitudes or beliefs.	4.21	0.42
	The session design effectively built student understandings of the CREATE process		
	The design of the session provided opportunities for "minds on" thinking about science content through primary literature		
Instruction of session	The implementer effectively incorporated instructional strategies appropriate for the purposes of the CREATE session and the needs of the learners	4.41	0.46
	Constructivism (students constructing new understandings on existing knowledge) was present throughout the session		
	Teacher role was that of facilitator rather than lecturer		
Nature of science	The nature of science was portrayed as presuming that things and events in the universe occur in consistent patterns that are comprehensible through careful, systematic study	3.90	0.49
	Students displayed abilities of inquiry (investigating, analyzing, explaining, evaluating, etc.)		
Science content in session	Science content was appropriate for the purpose of the CREATE session and the background of the students	4.28	0.58
	Appropriate connections were made from the science content to real world science contexts through the CREATE process		
	Relevant science concepts were explicitly addressed in the lesson to promote a deeper understanding of content		

Note: The OE used a five-point scale for rating the session using a modified Weiss protocol. 1, *not at all*; 5, *to a great extent*. The range of individual implementer (mean) scores for each category was design of session, 3.68–4.63; instruction of session, 3.78–4.73; nature of science, 3.30–4.51; science content in session, 3.39–4.78.

Her observations provided detailed information about the range of teaching methods and behaviors employed by each implementer.

Overall, the OE observational data confirmed that the faculty implementers applied CREATE pedagogy in alignment with workshop training. This outcome supports the efficacy of our workshop model. We note that many of the implementers had practiced teaching with CREATE during the workshop period. These mock class sessions led to robust discussions about potential challenges that might emerge when applying CREATE methods for the first time. Feedback from their workshop peers provided specific ideas for solutions or modifications for adapting their CREATE modules for their courses. After the workshop, we (the PIs) provided the implementers with significant feedback on their course syllabi and class plans during course preparation. Throughout the term, we maintained support through periodic conference calls and email communications. These interactions may have contributed to the fidelity of CREATE implementation. As was noted by Khatri and colleagues (2016), broad and sustained adoption of educational interventions often depends on the level of support provided to potential faculty users.

Finding 2: Positive student changes on cognitive and affective assessments. Across nine courses, the students demonstrated a small, significant gain (with an effect size [ES] of .78; table 4) on the test for experimental design ability (EDAT, Sirum and Humburg 2011). A comparison between pre- and postimplementation scores indicates a modest gain,

with fewer students scoring at the lowest end of the EDAT scale (supplemental figure S2). In addition, the students improved significantly when they were tested on their ability to critically evaluate claims on the basis of data (CTT). The pre- versus postcourse scores revealed a significant increase in logical reasoning (ES 1.79) concurrent with a significant decrease in illogical justifications (ES 1.47; table 4). Neither the EDAT nor the CTT was specific to the course's content; therefore, these changes reflect general skill development. Notably, these outcomes suggest that CREATE courses can facilitate the development of core competencies (e.g., the ability to apply the process of science), which is recommended by experts in biology education (AAAS 2015). More broadly, the transfer of critical thinking and science process skills is a desirable but difficult to achieve goal across STEM disciplines (NRC 2012, AAAS 2015). Students completing CREATE-based courses may be primed for future growth by enhanced skill development or practiced behaviors (critical thinking, self-regulated learning).

A frequent concern expressed by the workshop participants was balancing CREATE pedagogy with a perceived need for high content coverage, a universal issue facing STEM educators (Henderson and Dancy 2007, Coil et al. 2010, Andrews and Lemons 2015). We used closed-book concept map assessment as an indicator of the students' knowledge acquisition, retention, and application. The structural complexity of concept map elements (e.g., the increased number of concepts and associated links) has been associated with positive changes in content mastery (e.g., Andrews et al. 2008). We observed significant differences in student

Table 4. Paired sample *t*-tests of the mean difference (posttreatment – pretreatment score) of student outcomes on two cognitive assessments.

Instrument	Difference	2.5% confidence limit	97.5% confidence limit	*t*	*df*	*P*	Cohen's *d*
EDAT[a]	0.51	0.00	1.02	2.33	8	.048	0.78
CTT logical[b]	0.41	0.25	0.57	5.67	9	<.001	1.79
CTT illogical[b]	−0.27	−0.40	−0.14	4.66	9	.001	−1.47

[a]The assessment was scored for 165 students across nine institutions (2–10, see table 1).
[b]The assessment was scored for 176 students across all 10 implementations. For the CTT instrument, we subtracted each student's prescore from their postscore for each of the four questions. We then calculated an average (composite) difference in both logical and illogical statements for each student. To account for the lack of independence among students in the same course, we averaged across composite differences for the students within each class. A negative value indicates a decrease in illogical statements after the course.

Table 5. Paired sample *t*-tests of the mean difference (posttreatment – pretreatment score) of student outcomes on concept map assessment.

Category	Difference	2.5% confidence limit	97.5% confidence limit	*t*	*df*	*P*	Cohen's *d*
Number of concepts	2.99	0.47	5.50	2.81	7	.0262	0.99
Number of links	5.48	0.56	10.40	2.64	7	.0336	0.93
Number of labeled links	5.05	0.57	9.52	2.67	7	.0321	0.94
Number of flawed links[a]	−1.69	−3.60	0.23	2.80	3	.0676	1.40

Note: Eight of the 10 implementations included the concept map task, before and after the course (institutions 1, 3–8, 10). The number of paired student outcomes for the first three measurements was 161.
[a]The faculty implementers at four institutions assessed student concept maps for the number of flawed links. The faculty were blind to the designation as before or after. The number of paired student outcomeswas 66.

c-map outcomes (table 5). The postcourse maps were more complex (e.g., more concepts, links, and labeled links) across eight course implementations (figure 1, table 5). In a subset (four implementations) analyzed for flawed links, there was a marginal decrease in the number of such links (table 5; see the supplemental material for institutional variation). Overall, the outcomes of the c-map assessment suggest that CREATE pedagogy does not constrain an instructor's ability to focus on course content and, in fact, may offer creative methods to enhance student learning.

We also sought evidence of affective changes by assessing the students' attitudes, beliefs, and perceptions using the SAAB survey (Hoskins et al. 2011). Six of the subcategories probed the students' confidence across a range of science process skills related to reading primary scientific articles. The students demonstrated significant gains in self-assessed confidence across all six categories (ES > 1.0 for all; table 6). These changes suggest that CREATE courses promote self-efficacy with respect to science process skills, an area of increasing focus in STEM (Trujillo and Tanner 2014, Kenyon et al. 2016, Randler et al. 2016, Connell et al. 2016). As was described by Bandura (1977), self-efficacy reflects an internal measure of a student's confidence in the ability to complete or perform a goal-directed task within a particular domain. In CREATE courses, students apply and practice a diverse range of tasks as they deconstruct scientific studies and design new experiments. The assignments (e.g.,

constructing a concept map, annotating figures, transforming data) and in-class activities (e.g., a grant panel discussion and collegial arguments) may be considered mastery experiences that foster self-efficacy (Usher and Pajares 2008).

The SAAB survey also probed epistemological categories relevant to science. Epistemological beliefs are thought to play an important role in cognitive development, and some facets of learning may depend on shifts in those beliefs (Pintrich 2004). Epistemological beliefs also influence the development of reflective judgment and may affect students' use of particular cognitive approaches for evaluating and constructing knowledge (King and Kitchener 1994, Hofer and Pintrich 1997). We found changes reflecting positive shifts in three epistemological categories: the certainty of knowledge, the sense of scientists, and the sense of scientists' motives (table 6). The latter two categories suggest that the students acquired new understanding of scientists as individuals. An important element of CREATE pedagogy is fostering communication between students and scientists. The implementers sent (via email) student-generated questions regarding research life and researchers' personal experiences to the authors of articles analyzed in class. Class discussion of responses may have contributed to shifts in the students' views of scientists.

A change in their certainty of knowledge suggests that the students gained more nuanced views of science as a domain of knowledge that is dynamic and subject to continual

Table 6. Paired-sample *t*-tests of the mean difference (posttreatment – pretreatment score) in student scores on the affective SAAB survey.

Category	Difference	2.5% confidence limit	97.5% confidence limit	*t*	*df*	*P*	Cohen's *d*
Decoding primary literature	0.61	0.41	0.81	6.96	9	<.001	2.20
Interpreting data	0.28	0.13	0.44	4.08	9	.003	1.29
Active reading	0.52	0.37	0.67	7.99	9	<.001	2.52
Visualization	0.55	0.34	0.77	5.79	9	<.001	1.83
Thinking like scientist	0.47	0.34	0.59	8.47	9	<.001	2.68
Research in context	0.22	0.11	0.33	4.39	9	.002	1.39
Certainty of knowledge (R)	0.11	0.01	0.21	2.49	9	.035	0.78
Collaboration	0.08	−0.08	0.24	1.12	9	.292	0.35
Creativity	0.10	−0.04	0.23	1.67	9	.130	0.53
Innate ability (R)	0.07	−0.11	0.25	0.86	9	.414	0.27
Known outcomes (R)	−0.14	−0.34	0.05	1.71	9	.121	−0.54
Sense of motives	0.15	0.00	0.31	2.23	9	.052	0.70
Sense of scientists	0.56	0.36	0.77	6.18	9	<.001	1.95

Note: The bolded rows identify the six subcategories of the SAAB survey that address student beliefs and confidence in abilities required for processing and analyzing scientific literature. The other rows identify the seven subcategories addressing epistemological beliefs. The assessment was scored for 176 students. Abbreviation: R, reverse-scored statements.

Table 7. Regression coefficients (*b*), likelihood ratio test statistics, and marginal and conditional R^2 for each model of change in EDAT scores.

Explanatory variable	*b*	$b_{2.5\%}$	$b_{97.5\%}$	χ^2	*df*	*P*	R^2_M	R^2_C
Tenure status	−0.99	−1.56	−0.41	9.79	1	.002	.06	.06
Class size	−0.03	−0.07	0.00	3.52	1	.061	.03	.04
Teaching experience	−0.41	−1.05	0.24	1.45	1	.228	.01	.05
Selectivity	0.26	−0.42	0.94	0.56	1	.453	.01	.05

Note: *n* = 165 students.

revision. Similar gains have been found in other CREATE classes (first-years, Gottesman and Hoskins 2013; 4-year or upper level, Stevens and Hoskins 2014; 2-year, Kenyon et al. 2016). Shifting the epistemological beliefs of science students can be difficult, even in the context of undergraduate research experiences (Hunter et al. 2007). Many DBER researchers focus on course-based research experiences (CUREs) for addressing learning about the nature of science (Auchincloss et al. 2014, Ballen et al. 2017). CREATE methodology, which is explicitly focused on how scientific knowledge develops and changes through research, may be a worthy pedagogical alternative to achieve this goal, especially if resource-intensive CUREs are not an option for individual faculty or institutions.

Finding 3: Contextual factors that do and do not associate with student cognitive outcomes. Faculty characteristics such as professional status and teaching experience may influence instructional practices, including the application of

innovative methods. Untenured instructors or untenured tenure-track faculty may be hesitant to use new forms of instruction that are counter to department norms. Teaching experiences likely have an impact on teaching decisions (Andrews and Lemons 2015, Turpen et al. 2016, Walter et al. 2016). For example, individuals with significant experience may be less inclined to fully change tactics after workshop training and more inclined to revert to old habits (e.g., lecturing; Silverthorn et al. 2006). Faculty with less experience might be more likely to embrace innovative practices (Ebert-May et al. 2011). In the models we examined, years of teaching experience did not correlate with cognitive outcomes for students. The lack of association suggest that teaching experience is not an impediment to the effective use of CREATE pedagogy (as measured by positive student outcomes). In contrast, faculty status (untenured versus tenured) was related to student changes on a single measure, the EDAT instrument (table 7). The students in courses with tenured faculty demonstrated positive shifts, whereas
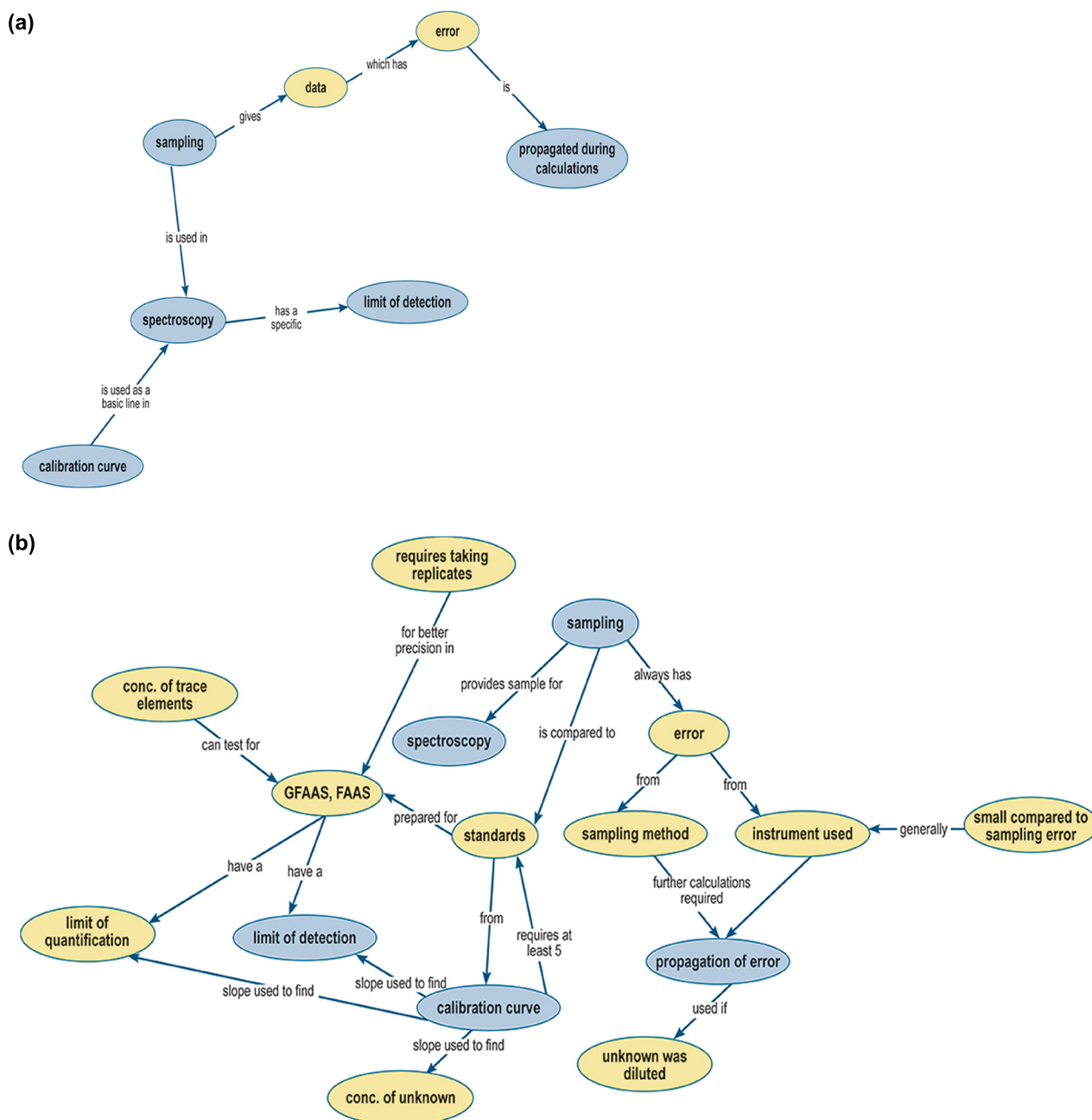
(a)

(b)

Figure 1. C-map assessment from one student (institution 10) (a) before and (b) after the course. The blue (dark) bubbles indicate seed terms selected by the instructor on the basis of the course's content. The yellow (light) bubbles identify concepts and terms added by the student.

Table 8. Regression coefficients (b), likelihood ratio test statistics, and marginal and conditional R^2 for each model of change in CTT logical scores.

Model	b	$b_{2.5\%}$	$b_{97.5\%}$	χ^2	df	P	R^2_M	R^2_C
Class size	−0.02	−0.03	−0.01	12.52	1	<.001	.12	.12
Selectivity	−0.09	−0.38	0.20	0.36	1	.547	.01	.09
Tenure status	−0.09	−0.41	0.23	0.27	1	.602	.00	.09
Teaching experience	0.02	−0.27	0.30	0.01	1	.916	.00	.09

Note: n = 176 students.
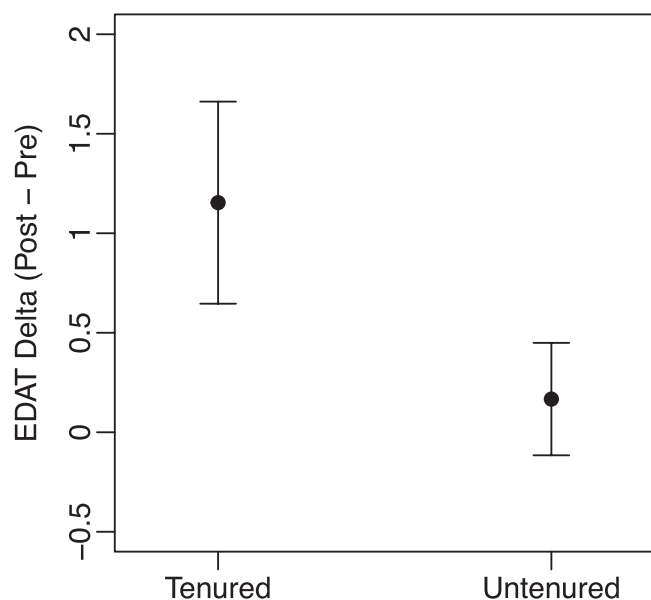
Figure 2. The relationship of change in EDAT scores to implementer tenure status. The error bars represent 95% confidence intervals.

Figure 3. The relationship of change in CTT logical scores with class size. The best-fit line is based on the parameter estimates in table 8.

those in courses taught by untenured instructors, on average, showed minimal change (figure 2). This association could not be explained by a difference in student cohorts based on the pretreatment scores on the EDAT (*t*-test, *p* > .05; see the data set on *http://datadryad.org*). We also note that there was not a simple relationship between experience and tenure status, because some untenured faculty had been teaching for more than 6 years. What may account for the association between student changes on the EDAT and faculty status? This outcome may reflect faculty differences due to status. For example, tenured faculty may have more flexibility in course selection or design (e.g., determining the types of formative and summative assessments), which may affect student outcomes. We acknowledge that the analysis may confound faculty status with course or student characteristics that were unknown to us (e.g., course topic, GPA, previous coursework, or ongoing research experiences). Therefore, we interpret this outcome with caution, especially given the small number of faculty involved (10). Future studies are clearly needed to address this potential relationship.

The difference scores on the CTT logical responses were negatively correlated to class size (table 8, figure 3), whereas changes in CTT illogical responses were not related to any of the variables under investigation (table S1). Increased variability among the students in larger classes may account for the former outcome. Alternatively, class size may influence the quantity and quality of student–faculty interactions, which could have an impact on individual skill development. Whether class size is a proxy for other factors that influence the development of critical thinking skills (e.g., writing ability) remains an open question. We did not have access to student information such as individual academic standing (e.g., GPA)
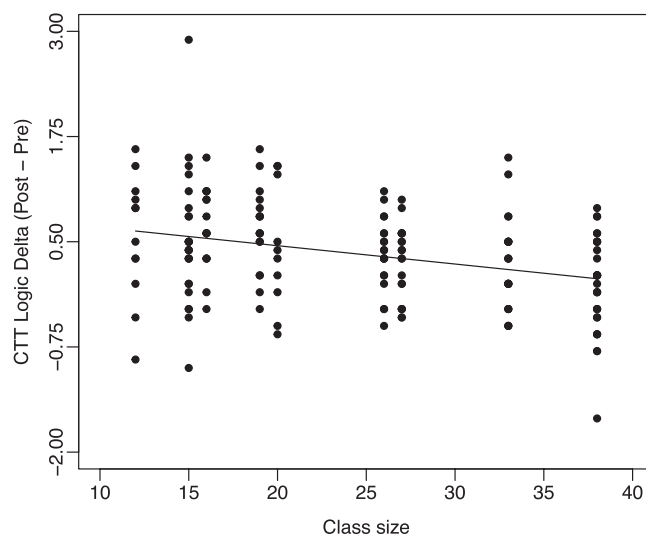
or course history. Instead, as a general indicator of student academic preparation, we used the Carnegie undergraduate profile classifications (*http://carnegieclassifications.iu.edu/ methodology/ugrad_profile.php*). We found no evidence that institutional differences in student populations were related to CREATE cognitive gains on the EDAT and CTT.

Conclusions

As more effort and resources are committed to improving STEM education, it is imperative that DBER researchers assess both the quality of faculty professional development experiences and the question of how such experiences affect downstream teaching practices and student learning (AAAS 2015, Derting et al. 2016). This study provides valuable insights into the downstream effects of CREATE workshops on faculty and students. The 10 faculty implementers taught at different types of institutions located throughout the United States, and they varied in their teaching experience and professional status. The courses ranged from introductory to upper level and spanned different disciplines and subdisciplines (e.g., ecology, genetics). Teaching observations by an experienced external evaluator indicated a high level of fidelity in applying CREATE methods across courses differing in size, type, and level. Moreover, the assessments revealed significant student gains in cognitive (e.g., experimental design, critical thinking) and affective (e.g., confidence in science process skills, epistemological maturation) domains in these new CREATE courses. The student gains found in this study replicate those from previous studies of CREATE implementations (4-year, Stevens and Hoskins 2014; 2-year, Kenyon et al. 2016). As was noted by Makel and Plucker (2014), "If education research is to be relied on to develop sound policy and practice, then conducting replications is essential to moving toward a more reliable and

trustworthy understanding of educational environments." Our findings support the idea that CREATE workshop training translates successfully to the classroom while also identifying situational factors that may influence implementation and its impact on student learning. Overall, this research provides a model for dissemination and evaluation of pedagogical innovations.

CREATE curricular resources (e.g., roadmaps) produced from the CREATE faculty development workshops can be found at *www.teachcreate.org*.

Acknowledgments

Supplemental material

Supplemental data are available at *BIOSCI* online.

References cited

[AAAS] American Association for the Advancement of Science. 2015. Vision and Change in Undergraduate Biology Education: Chronicling Change, Inspiring the Future. AAAS, National Science Foundation. http://visionandchange.org/chronicling-change.

Amrein-Beardsley A, Popp SE. 2012. Peer observations among faculty in a college of education: Investigating the summative and formative uses of the Reformed Teaching Observation Protocol (RTOP). Educational Assessment, Evaluation, and Accountability 24: 5–24.

Andrews TC, Lemons PP. 2015. It's personal: Biology instructors prioritize personal evidence over empirical evidence in teaching decisions. CBE Life Sciences Education 14: ar7.

Andrews KE, Tressler KD, Mintzes JJ. 2008. Assessing environmental understanding: An application of the concept mapping strategy. Environmental Education Research. 14: 519–536.

Andrews TM, Leonard M, Colgrove CA, Kalinowski ST. 2011. Active learning not associated with student learning in a random sample of college biology courses. CBE Life Sciences Education 10: 394–405.

Auchincloss LC, Laursen SL, Branchaw JL, Eagan K, Graham M, Hanauer DI, Lawrie G, McLinn CM, Pelaez N, Rowland S, Towns M, Trautmann NM, Varma-Nelson P, Weston TJ, Dolan EL. 2014. Assessment of course-based undergraduate research experiences: A meeting report. CBE Life Sciences Education 13: 29–40.

Ballen CJ, Blum JE, Brownell S, Hebert S, Hewlett J, Klein JR, McDonald EA, Monti DL, Nold SC, Slemmons KE, Soneral PAG, Cotner S. 2017. A call to develop course-based undergraduate research experiences (CUREs) for nonmajors courses. CBE Life Science Education 16: mr2.

Bandura, A. 1977. Self-efficacy: Toward a unifying theory of behavior change. Psychological Review 84: 191–215.

Banilower ER, Heck DJ, Weiss IR. 2007. Can professional development make the vision of the standards a reality? The impact of the national science foundation's local systemic change through teacher enhancement initiative. Journal of Research in Science Teaching 44: 375–395.

Best JW, Kahn JV. 2006. Descriptive data analysis. Pages 353–401 in Research in Education, 10th ed. Pearson Education.

Coil D, Wenderoth MP, Cunningham M, Dirks C. 2010. Teaching the process of science: Faculty perceptions and an effective methodology. CBE Life Sciences Education 9: 524–535.

Connell GL, Donovan DA, Chambers TG. 2016. Increasing the use of student-centered pedagogies from moderate to high improves student learning and attitudes about biology. CBE Life Sciences Education 15: ar3

Dancy M, Henderson C, Turpen C. 2016. How faculty learn about and implement research-based instructional strategies: The case of peer instruction. Physical Review Physics Education Research 12: 010110.

Derting TL, Ebert-May D. 2010. Learner-centered inquiry in undergraduate biology: Positive relationships with long-term student achievement. CBE Life Sciences Education 9: 462–472.

Derting TL, Ebert-May D, Henkel TP, Maher JM, Arnold B, Passmore HA. 2016. Assessing faculty professional development in STEM higher education: Sustainability of outcomes. Science Advances 2 (art. e1501422).

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

Ebert-May D, Derting TL, Henkel TP, Middlemis Maher J, Momsen JL, Arnold B, Passmore HA. 2015. Breaking the cycle: Future faculty begin teaching with learner-centered strategies after professional development. CBE Life Sciences Education 14 (art. ar22).

Freeman S, Eddy SL, McDonough M, Smith MK, Okoroafor N, Jordt H, Wenderoth MP. 2014. Active learning increases student performance in science, engineering, and mathematics. Proceedings of the National Academy of Sciences 111: 8410–8415.

Gottesman AJ, Hoskins SG. 2013. CREATE cornerstone: Introduction to scientific thinking, a new course for STEM-interested freshmen, demystifies scientific thinking through analysis of scientific literature. CBE Life Sciences Education 12: 59–72.

Henderson C, Dancy MH. 2007. Barriers to the use of research-based instructional strategies: The influence of both individual and situational characteristics. Physical Review Special Topics: Physics Education Research 3: 020102.

Henderson C, Beach A, Finkelstein N. 2011. Facilitating change in undergraduate stem instructional practices: An analytic review of the literature. Journal of Research in Science Teaching 48: 952–984.

Henderson C, Dancy M, Niewiadomska-Bugaj M. 2012. Use of research-based instructional strategies in introductory physics: Where do faculty leave the innovation-decision process? Physical Review Special Topics: Physics Education Research 8: 020104.

Hofer BK, Pintrich PR. 1997. The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning. Review of educational research. 67: 88–140.

Hoskins SG, Krufka A. 2015. The CREATE strategy benefits students and is a natural fit for faculty. Microbe 10: 108–112.

Hoskins SG, Stevens LM, Nehm RH. 2007. Selective use of the primary literature transforms the classroom into a virtual laboratory. Genetics 176: 1381–1389.

Hoskins SG, Lopatto D, Stevens LM. 2011. The CREATE approach to primary literature shifts undergraduates' self-assessed ability to read and analyze journal articles, attitudes about science, and epistemological beliefs. CBE Life Sciences Education 10: 368–378.

Hoskins S, Gottesman A, Kenyon K. 2017. CREATE two-year/four-year faculty workshops: A focus on practice, reflection, and novel curricular design leads to diverse gains for faculty at two-year and four-year institutions. Journal of Microbiology and Biology Education 18. doi:10.1128/jmbe.v18i3.1365.

Hunter AB, Laursen S, Seymour E. 2007. Becoming a scientist: The role of undergraduate research in students' cognitive, personal, and professional development. Science Education. 91: 36–174.

Kenyon KL, Onorato M, Gottesman AJ, Hoque J, Hoskins SG. 2016. Testing CREATE at community colleges: An examination of faculty perspectives and diverse student gains. CBE Life Sciences Education 15: ar8.

Khatri R, Henderson C, Cole R, Froyd JE, Friedrichsen D, Stanford C. 2016. Designing for sustained adoption: A model of developing

educational innovations for successful propagation. Physical Review Physics Education Research 12: 010112.

King PM, Kitchener KS. 1994. Developing Reflective Judgment: Understanding and Promoting Intellectual Growth and Critical Thinking in Adolescents and Adults. Jossey-Bass.

Love Stowell SM, Churchill AC, Hund AK, Kelsey KC, Redmond MD, Seiter SA, Barger NN. 2015. Transforming graduate training in STEM education. Bulletin of the Ecological Society of America 96: 317–323.

Lund TJ, Pilarz M, Velasco JB, Chakraverty D, Rosploch K, Undersander M, Stains M. 2015. The best of both worlds: Building on the COPUS and RTOP observation protocols to easily and reliably measure various levels of reformed instructional practice. CBE Life Sciences Education 14: ar18.

Makel MC, Plucker JA. 2014. Facts are more important than novelty: Replication in the education sciences. Educational Researcher 43: 304–316.

National Research Council. 2012. Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century. National Academies Press. https://doi.org/10.17226/13398.

Novak JD. 2003. The promise of new ideas and new technology for improving teaching and learning. Cell Biology Education 2: 122–132.

Pfund C, Miller S, Brenner K, Bruns P, Chang A, Ebert-May D, Fagen AP, Gentile J, Gossens S, Khan IM, Labov JB, Pribbenow CM, Susman M, Tong L, Wright R, Yuan RT, Wood WB, Handelsman J. 2009. Professional development. Summer institute to improve university science teaching. Science 324: 470–471.

Pintrich PR. 2004. A conceptual framework for assessing motivation and self-regulated learning in college students. Educational Psychology Review 16: 385–407.

Quillin K, Thomas S. 2015. Drawing-to-learn: A framework for using drawings to promote model-based reasoning in biology. CBE Life Sciences Education 14 (art. Es2).

R Core Team. 2016. R: A language and environment for statistical computing. R Foundation For Statistical Computing. (www.r-project.org)

Randler C, Demirhan E, Wüst-Ackermann P, Desch IH. 2016. Influence of a dissection video clip on anxiety, affect, and self-efficacy in educational dissection: A treatment study. CBE Life Sciences Education 15: ar1.

Rogers EM. 1995. Diffusion of Innovations, 4th ed. Free Press.

Sawada D, Piburn MD, Judson E, Turley J, Falconer K, Benford R, Bloom I. 2002. Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. School Science and Mathematics 102: 245–253.

Silverthorn DU, Thorn PM, Svinicki MD. 2006. It's difficult to change the way we teach: Lessons from the Integrative Themes in Physiology curriculum module project. Advances in Physiology Education 30: 204–214.

Sirum K, Humburg J. 2011. The experimental design ability test (EDAT) Bioscene. Journal of College Biology Teaching 37: 8–16.

Smith MK, Jones FH, Gilbert SL, Wieman CE. 2013. The Classroom Observation Protocol for Undergraduate STEM (COPUS): A new instrument to characterize university STEM classroom practices. CBE Life Sciences Education 12: 618–627.

Smith MK, Vinson EL, Smith JA, Lewin JD, Stetzer MR. 2014. A campus-wide study of STEM courses: New perspectives on teaching practices and perceptions. CBE Life Sciences Education 13: 624–635.

Stains M, Vickrey T. 2017. Fidelity of implementation: An overlooked yet critical construct to establish effectiveness of evidence-based instructional practices. CBE Life Sciences Education 16: p.rm1.

Stevens LM, Hoskins SG. 2014. The CREATE strategy for intensive analysis of primary literature can be used effectively by newly trained faculty to produce multiple gains in diverse students. CBE Life Sciences Education 13: 224–242.

Tanner K, Allen D. 2006. Approaches to biology teaching and learning: On integrating pedagogical training into the graduate experiences of future science faculty. CBE Life Sciences Education 5: 1–6.

Trujillo G, Tanner KD. 2014. Considering the role of affect in learning: Monitoring students' self-efficacy, sense of belonging, and science identity. CBE Life Sciences Education 13: 6–15.

Turpen C, Dancy M, Henderson C. 2016. Perceived affordances and constraints regarding instructors' use of peer instruction: Implications for promoting instructional change. Physical Review Physics Education Research 12: 010116.

Usher EL, Pajares F. 2008. Sources of self-efficacy in school: Critical review of the literature and future directions. Review of Educational Research 78: 751–796.

Van Vliet EA, Winnips JC, Brouwer N. 2015. Flipped-class pedagogy enhances student metacognition and collaborative-learning strategies in higher education but effect does not persist. CBE Life Sciences Education 14: p.ar26.

Walter EM, Henderson CR, Beach AL, Williams CT. 2016. Introducing the Postsecondary Instructional Practices Survey (PIPS): A concise, interdisciplinary, and easy-to-score survey. CBE Life Sciences Education 15: ar53.

Wieman C, Gilbert S. 2014. The Teaching Practices Inventory: A new tool for characterizing college and university teaching in mathematics and science. CBE Life Sciences Education 13: 552–569.

Yager RE. 2013. Exemplary College Science Teaching. NSTA Press.

Yelon SL, Kevin Ford J, Bhatia, S. 2014. How trainees transfer what they have learned: Toward a taxonomy of use. Performance Improvement Quarterly 27: 27–52.

Kristy L. Kenyon is a professor of biology and Bradley J. Cosentino is an associate professor of biology at Hobart and William Smith Colleges, in Geneva, New York. Alan J. Gottesman is a research assistant and Sally G. Hoskins is a professor of biology at City College of the City University of New York. Morgan E. Onorato (Hobart and William Smith) and Jamila Hoque (City College of the City University of New York) were undergraduate research assistants during this project.