Integrating Concepts in Biology Textbook Increases Learning: Assessment Triangulation Using Concept Inventory, Card Sorting, and MCAT Instruments, Followed by Longitudinal Tracking

Douglas B. Luckie,[†‡§*] Anne-Marie Hoskinson,[‖] Caleigh E. Griffin,[†‡] Andrea L. Hess,[†‡] Katrina J. Price,[†‡] Alex Tawa,[†‡] and Samantha M. Thacker[†‡]

[†]Lyman Briggs College, [‡]STEM Learning Laboratory, and [§]Department of Physiology, Michigan State University, East Lansing, MI 48825; [‖]Department of Biology and Microbiology, South Dakota State University, Brookings, SD 57007

ABSTRACT

The purpose of this study was to examine the educational impact of an intervention, the inquiry-focused textbook *Integrating Concepts in Biology* (*ICB*), when used in a yearlong introductory biology course sequence. Student learning was evaluated using three published instruments: 1) The Biology Concept Inventory probed depth of student mastery of fundamental concepts in organismal and cellular topics when confronting misconceptions as distractors. *ICB* students had higher gains in all six topic categories (+43% vs. peers overall, $p < 0.01$). 2) The Biology Card Sorting Task assessed whether students organized biological ideas more superficially, as novices do, or based on deeper concepts, like experts. The frequency with which *ICB* students connected deep-concept pairs, or triplets, was similar to peers; but deep understanding of structure/function was much higher (for pairs: 77% vs. 25%, $p < 0.01$). 3) A content-focused Medical College Admission Test (MCAT) posttest compared *ICB* student content knowledge with that of peers from 15 prior years. Historically, MCAT performance for each semester ranged from 53% to 64%; the *ICB* cohort scored 62%, in the top quintile. Longitudinal tracking in five upper-level science courses the following year found *ICB* students outperformed peers in physiology (85% vs. 80%, $p < 0.01$).

INTRODUCTION

In most settings, biologists can no longer limit themselves to pursuing only molecular or organismal methods, nor can they avoid using quantitative and interdisciplinary approaches (National Research Council [NRC], 2003; Association of American Medical Colleges and the Howard Hughes Medical Institute [AAMC-HHMI], 2009; American Association for the Advancement of Science [AAAS], 2011; Waldrop and Miller, 2015). For example, to understand large, rapidly changing ecosystems, biologists must be able to study long-term ecological research plots in the alpine tundra; read DNA gels; and use modern mathematical, statistical, computational, and technological tools. As a result, biology instruction and scholarly instruction at all levels must keep pace with these changes in the practice of research (AAAS, 2011; NRC, 2012, 2014; Next Generation Science Standards Lead States, 2013). A new textbook, *Integrating Concepts in Biology* (*ICB*; Campbell *et al.*, 2014), was designed to confront this "new normal" and enable instructors to engage students in regular practice of scientific inquiry inside the lecture room (Barsoum *et al.*, 2013; Campbell *et al.*, 2014; Wagner *et al.*, 2015).

The purpose of this research study was to look for evidence of impact of a single intervention, the *ICB* textbook, when adopted for a yearlong introductory biology

Kathryn E. Perez, *Monitoring Editor*

Submitted June 24, 2016; Revised January 25, 2017; Accepted January 30, 2017

CBE Life Sci Educ June 1, 2017 16:ar20

DOI:10.1187/cbe.16-06-0204

*Address correspondence to: Douglas B. Luckie (luckie@msu.edu).

course sequence already practicing reformed pedagogies (Reformed Teaching Observation Protocol [RTOP] levels III and IV; Sawada *et al.*, 2002; Ebert-May *et al.*, 2011). The *ICB* textbook rigorously implements recommendations and practices as described in *Vision and Change* (AAAS, 2011). While traditional textbooks often place content at the center and include scientific practice in the margins, the *ICB* textbook reverses that approach, and makes engaging in science practice central for students (Barsoum *et al.*, 2013; Prestwich and Sheehy, 2015).

We hypothesized the *ICB* curriculum could boost conceptual expertise and longitudinal performance but perhaps negatively impact short-term gains in rote content knowledge. Hence, during the yearlong intervention, data were collected using three published instruments. The Biology Concept Inventory (BCI; Klymkowsky and Garvin-Doxas, 2008) and Biology Card Sorting Task (BCST; Smith *et al.*, 2013) were used to detect changes in expertise, and a Medical College Assessment Test (MCAT) instrument (Luckie *et al.*, 2004) was used to assess content knowledge, as well as longitudinal cohort analysis to follow student performance in upper-level courses.

The BCI was selected because it probes depth of student mastery of fundamental concepts in both organismal and cellular topics when confronting strong distractors based on established frequent misconceptions (Klymkowsky and Garvin-Doxas, 2008; Klymkowsky *et al.*, 2010). Concept inventories are carefully developed over years. Students are initially interviewed, and their verbal responses are transcribed (Smith *et al.*, 2008). Later, additional students might respond to the same questions with extended-response or essay answers (Adams and Wieman, 2011). Common misconceptions that are held by students are slowly developed into multiple-choice answers as distractors to accompany a valid response (Anderson *et al.*, 2002; Smith and Tanner, 2010). Years of testing allow the researchers to evaluate the terminology and refine the wording by incorporating student vocabulary until each question is both valid (successful in communicating the true question meaning) and found to be reliable (repeatedly able to evaluate student understanding; Adams and Wieman, 2011). The BCI was used to detect whether or not there are deficiencies in student expertise (Garvin-Doxas and Klymkowsky, 2008; Klymkowsky and Garvin-Doxas, 2008). We predicted the BCI might indicate that students would show a deeper understanding of basic biological concepts due to the *ICB* curriculum. While a number of inventories exist for biology, most are quite narrow, focusing on topics such as natural selection (Anderson *et al.*, 2002), genetics (Smith *et al.*, 2008), or the central dogma (Newman *et al.*, 2016). The BCI is unique in that it is a broad concept inventory, spanning molecular, cellular, and organismal topics, and was therefore more appropriate to our study, which spanned topics introduced in yearlong introductory biology courses.

A second approach was also adopted to collect additional evidence regarding whether the *ICB* curriculum could boost conceptual expertise. The BCST was designed to assess whether students organize biological ideas superficially, as novices tend to do, or based on deep concepts (e.g., evolution, energy and matter), as experts do (Smith *et al.*, 2013). The way in which individuals organize subject-specific information is often an accurate indicator of developing expertise (Newell and Simon, 1972; Chi *et al.*, 1981). Chi and colleagues (1981) created and validated a problem-sorting approach for physics problems, and Smith and colleagues subsequently developed and validated an instrument using biology problems (Smith *et al.*, 2013). While assessing introductory biology students using the BCST, Smith and colleagues (2013) found that their data supported the prediction that novices would categorize problems based on surface features, rather than deep features or key concepts used by experts (Smith and Good, 1984). Our study's BCST data were collected by Hoskinson, Ebert-May, and colleagues in their study of 16 introductory biology courses, including *ICB* students, at our university (Hoskinson *et al.*, 2017).

We also hypothesized that the *ICB* curriculum, which is not as explicitly content focused as traditional textbooks, might negatively impact short-term gains in content knowledge. Since 2000, during the final week of each semester, students in our college have been given a content-based assessment instrument constructed from MCAT questions to assess individual performance (Luckie *et al.*, 2004). It was originally developed with the rationale that our science, technology, engineering, and mathematics (STEM) colleagues recognize the MCAT as an important instrument, which a Bloom evaluation also found respectable (Donnon *et al.*, 2007; Zheng *et al.*, 2008). Thus, a 40-question MCAT posttest was used as a content assessment in this study enabling a comparison of *ICB* student performance with the historical performance of students over the prior 15 years using traditional textbooks.

A strong foundation gained in introductory biology can lead to success in upper-level STEM courses and beyond (White and Arzi, 2005; Derting and Ebert-May, 2010). By specifically tracking student performance, one can test predictions of long-term effects of an intervention (Oakes, 1992; Helldén, 2005; Jeffreys, 2007; Wai *et al.*, 2010). We hypothesized the *ICB* curriculum could boost longitudinal performance. Thus, longitudinal tracking was used to look for later success (Voorhees and Lee, 2009; Creech and Sweeder, 2012). We performed longitudinal tracking of students who had completed the *ICB* curriculum and were entering into five upper-level STEM courses the following year: Physiology (PSL), Advanced Physiology I (Adv. PSL), Biochemistry (BCH), Advanced Biochemistry I (Adv. BCH), and Physics I (PHY). Within the same upper-level classroom settings, we compared the performance of *ICB* students with that of their peers who completed non-*ICB* introductory biology courses.

Our findings support those of Barsoum *et al.* (2013), who noted that performance of the *ICB* cohort surpassed peers at the end-of-year time point, and suggest the *ICB* approach may enable learning gains beyond those found using traditional content-focused textbooks, even in courses already using reformed pedagogies.

METHODS

Participants

With the approval of the Institutional Review Board, data were collected from all students who completed the yearlong *ICB* course and control students from Michigan State University. The MSU registrar's office provided ACT performance data and all students were enrolled at MSU and participant consent obtained. The yearlong *ICB* course sequence, when offered, was Introductory Biology I (LB144, sections 1–6, 109 students) in Fall 2014 and Introductory Biology II (LB145, sections 1–5, 89 students), offered in Spring 2015. BCI control participants

were enrolled in an equal-sized section of the same course offered in the same classroom during the same semesters. The instructor used extensive and regular modern reformed pedagogies with a traditional textbook (Freeman, 2010, *Biological Science*, 4th Edition) in the control courses: Introductory Biology I (LB144, sections 7–12, 117 students), offered in Fall 2014, and Introductory Biology II (LB145, sections 6–10, 96 students), offered in Spring 2015. BCST control participants in this study were students enrolled in either their first or second course in introductory biology, during Spring 2014, Fall 2014, or Spring 2015. The 751 control participants were enrolled in one of 16 different course sections. These instructors used a wide range of active and passive pedagogies and traditional textbooks (Freeman *et al.*, 2014, *Biological Science*, 5th edition; Raven *et al.*, 2011, *Biology*, 9th edition; and Reece *et al.,* 2013, *Campbell Biology*, 10th edition). There were 2164 students who were participants in the longitudinal cohort analysis. These were students enrolled in five courses during Fall 2015 and Spring 2016 (PSL $n = 943$, Adv. PSL $n = 207$, BCH $n = 738$, Adv. BCH $n = 146$, PHY $n = 130$).

ICB Course Pedagogy

The lecture meetings in the *ICB* course were best described as minilectures separated by regular use of active and cooperative think–pair–share types of exercises facilitated by use of clickers. Students had frequent homework assignments, which were tightly tied to explicit use of the textbook only. *Integrating Concepts in Biology* by Campbell, Heyer, and Paradise (Campbell *et al*., 2014) was used as an online textbook. The *ICB* course included homework provided with the LON-CAPA (CourseWeaver) and TopHat courseware platforms. This enabled *ICB* students to pause during active reading and provide extended-response answers to the "Integrating Questions" embedded in the *ICB* textbook. Their responses were evaluated by human graders and scores were recorded. This layer of online technology enabled use of logs in retrospective tracking of the path each student chose to pursue throughout the course.

Biology Concept Inventory

The BCI is a diagnostic tool developed using traditional methods required for a concept inventory (Garvin-Doxas and Klymkowsky, 2008; Klymkowsky and Garvin-Doxas, 2008; Klymkowsky *et al*., 2010). The BCI is a valid and reliable multiple-choice instrument, available online. It consists of 30 questions spanning six biological categories with distractors based on established misconceptions gleaned from subject interviews. The BCI's categories and corresponding questions are: diffusion and drift (questions 1, 5, 25, 29, 30), energetics and interactions (Q2, 3, 17, 18), molecular properties and functions (Q10, 11, 13, 19, 20, 27), genetic behaviors (Q7, 15, 16, 21, 22, 24, 28), evolutionary mechanisms (Q4, 6, 12, 14, 26), experimental design (Q8, 9). During this study, *ICB* student performance on the concept inventory was compared with performance of peers enrolled in another equal-sized section of the same introductory biology course with a different instructor using the same reformed pedagogies but a traditional content-focused textbook. Students were not randomly assigned but were unaware of the identities of instructors when selecting course sections. The pretest was administered at the beginning of Introductory Biology I courses in Fall semester 2014, the midtest at end of the Fall semester 2014, and the posttest at the end of Introductory Biology II in Spring 2015.

Biology Card Sorting Task

The BCST is an instrument developed by Smith and colleagues and is designed to measure students' biology expertise (Smith *et al.*, 2013). The BCST instrument was an adaptation of a card-sorting tool originally developed for physics students (Chi *et al.*, 1981). Early work done by Reif (Larkin and Reif, 1979) indicated that subjects initially distinguish a problem based on abstract concepts associated with a specific "problem schemata." These frameworks are often not consciously apparent, even to those considered to be experts (Dreyfus and Dreyfus, 2005). Problem sorting is an elegant instrument that can quickly differentiate novices from experts based on the well-documented principle that novices tend to use superficial traits to organize ideas, whereas experts use deep principles. Problem sorting has long been used in cognitive psychology to understand how people form and connect concepts. Biology problems were extracted from introductory college biology textbooks. Each problem was chosen so that it included one and only one superficial trait (here, organisms: humans, plants, animals, or microbes) and one and only one deep concept (here, core concepts from *Vision and Change*: energy and matter, structure and function, information storage, and evolution). Students read each problem and were then directed to sort (group) the problems together under one of two sorting conditions. Some students were provided the core concepts, to evaluate whether they were able to associate problems with core concepts. Other students were asked to simply sort the problems in ways that made sense to them, to explore their conceptual frameworks.

The BCST was part of a larger investigation of 16 introductory biology courses (Hoskinson *et al.*, 2017), wherein the methods used for the study described in this article are provided in greater detail. Performance by peers enrolled in a number of comparable introductory biology courses using traditionally content-focused textbooks was used for BCST controls. Participants in this investigation were asked to sort 16 college-level biology problems under the following categories: evolution by natural selection, pathways and transformations of energy and matter, storage and passage of information, structure and function (Smith *et al.*, 2013). Each of the 16 problems contained a single deep feature and a single surface/superficial feature (plant, insect, human, microorganism). When a subject placed two problems in the same hypothesized deep-feature category, this was recorded as a "deep pair." When a subject placed two cards in the same superficial-feature category, this was recorded as a "superficial pair." Three cards placed in the same deep-concept category resulted in a "deep triplet." The maximum correct number of deep problem pairs was 24, and triplets were 16. Normalization of gain (change within context of headroom) was performed as described (Hoskinson *et al.*, 2017). Like the BCI, the BCST was administered at the beginning of the academic year, at the end of the first semester (midpoint), and finally, at the end of the year. At the pre-, mid-, and posttest time points, *ICB* students' performance on the BCST was compared with the performance of peers enrolled in other introductory biology courses using traditional content-focused textbooks.

MCAT Instrument (MAT)

The Medical Assessment Test (MAT) is a small, standardized content exam that has been used historically as a regular posttest for Introductory Biology II students since 2000. The MAT is composed of MCAT study questions developed, validated, and purchased from the Association of American Medical Colleges (Luckie *et al.*, 2004, 2012, 2013). The MAT exam is a 40-question, multiple-choice test composed of relevant passage–style questions. MCAT passage questions have been studied by others and deemed to assess higher-level content knowledge than typical multiple-choice exams (Zheng *et al.*, 2008). The MAT consisted of questions from five general topic categories: cell structure and function, oncogenes/cancer, cellular respiration, microbiology, and DNA structure and function. Performance of each individual student on the MAT as a whole and on questions related to each category was examined, along with the performance of historical control students. The MAT instrument used in this study is provided online (Luckie *et al.*, 2004).

Longitudinal Cohort Analysis

Student performance was tracked in five upper-level science courses during the following academic year (2015–2016). Grades earned by the cohort of all *ICB* students were compared with those earned by other MSU students who enrolled in upper-level science courses but did not take the *ICB* introductory course. The upper-level courses examined were Physiology (PSL), Advanced Physiology I (Adv. PSL), Biochemistry (BCH), Advanced Biochemistry I (Adv. BCH) and Physics I (PHY). RTOP performance and differences in pedagogies used by upper-level instructors were not assessed or controlled, yet *ICB* and peer-control students shared the same experience in those learning environments. The data consisted of grades (final total points earned) in these upper-level science courses and data were analyzed in several ways. First, we compared the entire *ICB* student cohort with its peer cohort in each upper-level course. Second, much like investigators in a drug study, we were curious about the effect of low or high dosage of the *ICB* intervention on human subjects. Hence, to identify "low-dosage" subjects, we used computer server logs to explore post hoc student usage data. We identified 1) students who never purchased the *ICB* textbook, 2) students who never used the online textbook, and 3) students who did not answer online "Integrating Questions" embedded in the textbook readings (all "Integrating Questions" in homework were extended-response type and were evaluated by human graders). In addition, the cohorts of *ICB* students who enrolled in an *ICB* course only for a single semester were tracked and compared with peers.

Statistical Evaluation

Data from instruments and tracking were normalized for variations in each cohort's prior academic performance using ACT scores (first with ACT science score, secondarily using ACT composite score; Hake, 1998) unless otherwise indicated. Microsoft Excel was used to generate charts and box plots, organize the data sets, and perform statistical tests. Student's two-tailed *t* test results (*p* values) are those listed for all figures. Figure legends indicate trial numbers, and error bars on figures were generated by calculating the standard error of the mean (SEM) unless otherwise indicated.

RESULTS

BCI: *ICB* Students Performed Better Than Peers at the Posttest Stage

At the pre-, mid- and posttest time points, *ICB* students performance on the concept inventory was compared with performance of peers enrolled in another equal-sized section of the same introductory biology course that was using the same reformed pedagogies but a traditional content-focused textbook. Overall, students who participated in the *ICB* curriculum had significantly higher gains at the end of the academic year compared with students in the traditional biology course ($+43.19 \pm 7.02\%$, $p < 0.01$; Figure 1A). At the posttest stage, *ICB* students had a greater percentage increase in all concept inventory categories (Figure 1A). For *ICB* students, the greatest gains were found in the Diffusion and Drift category ($+12.56 \pm 5.10\%$) and the Experimental Design category ($+20.25 \pm 4.09\%$). The results suggest that greatest gains manifest or become detectable during the second half of the *ICB* sequence, perhaps due to an additive effect of two semesters.

BCST: *ICB* Students Achieved Highest Scores in Deep Triple Sort

As was done with the concept inventory, the Biology Card Sorting Task was administered at the pre-, mid- and posttest time points, but in this case, the controls were students enrolled in other introductory biology courses, all of which used traditional textbooks. Deep pair sorting occurred when a student paired two problems like experts, while a deep triplet sort occurred when a student grouped three problems (of four possible) like an expert. At the posttest stage, *ICB* students categorized problems like an expert precisely as well as control peers (Figure 2). After normalization, no significant difference was found in pre- to posttest gains overall, except that a stronger understanding was seen for *ICB* students in the subcategory of structure and function for both pair and triplet sorts (pair: $77 \pm 10\%$ vs. $28 \pm 3\%$, $p < 0.01$), and a weaker understanding of evolution for triplet sort data (triplet: $-6.9 \pm 1.7\%$ vs. $26 \pm 6.5\%$, $p < 0.05$; Figure 2). The low trial number for the *ICB* students in the triplet analysis ($n = 21$) limits robust interpretation.

MAT: *ICB* Students Scored in the Top Quintile

The normalized MAT scores from years 2000–2013 ranged from 53.39 to 64.21%. *ICB* students scored a 62.22%, thus within the top quintile (81.6%) of the highest performance in 15 years (Figure 3A). *ICB* student performance ($62.22 \pm 1.43\%$) was statistically greater than historical cohorts from years 2000–2001 ($53.39 \pm 1.96\%$, $p < 0.05$) and lower than the highest historical MCAT score achieved in 2011 ($64.21 \pm 1.84\%$; Figure 3A and inset). When MCAT subtopics were examined, *ICB* students performed within historical norms (Figure 3B), and again, statistical separation was only seen versus the 2000–2001 and 2011 cohorts.

Longitudinal Analysis: *ICB* Students Outperformed Peers in Upper-Level Physiology

The following year, students from the *ICB* cohort were tracked into five upper-level STEM courses: Physiology (PSL, Adv. PSL), Biochemistry (BCH, Adv. BCH), and Physics (PHY). In PSL, *ICB*

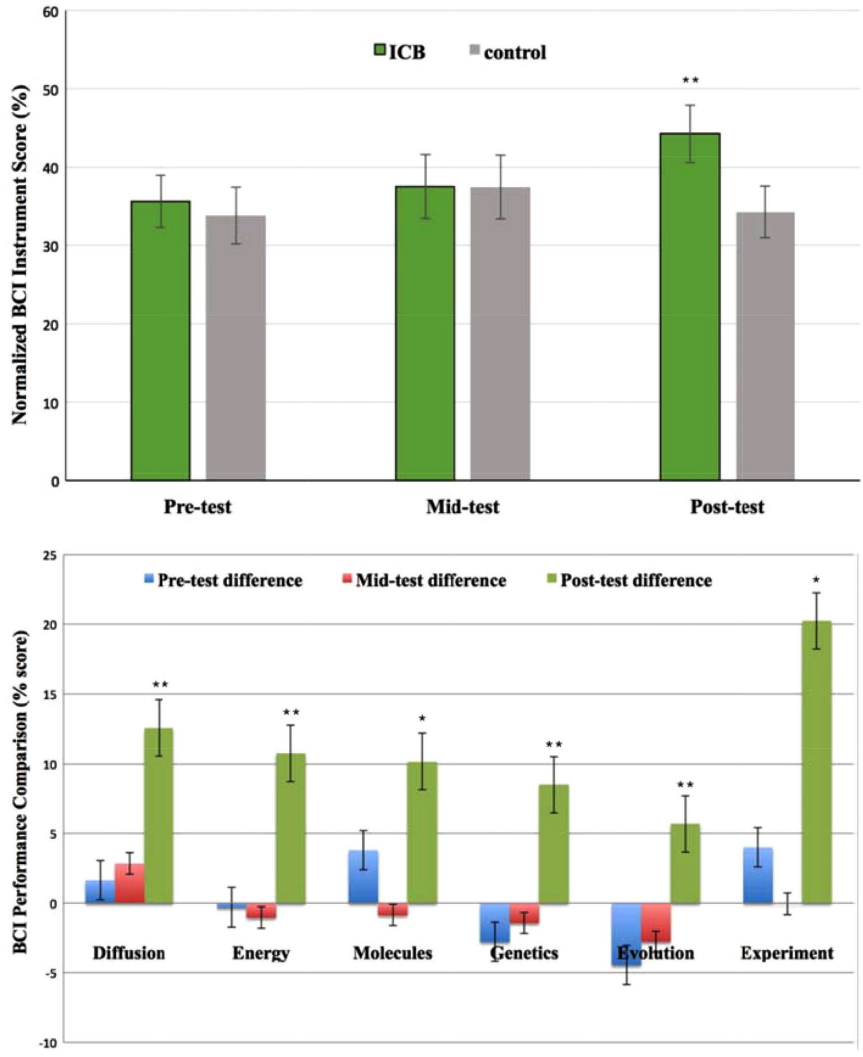

FIGURE 1. BCI performance of *ICB* students and controls. (A) Performance of *ICB* students and control cohort at the pre-, mid-, and posttest stages. (B) Percent difference of *ICB* vs. control for each subcategory topic tested at each test stage. *ICB* students ($n = 76$) and controls ($n = 98$). Error bars = SEM; *$p < 0.05$; **$p < 0.01$.

students outperformed peers (85.71 ± 8.6% vs. 80.71 ± 13.64%, $p = 0.021$), but no significant difference was found in other courses (Figure 4A). A set of students who completed only the first semester of the yearlong *ICB* curriculum were also tracked as they moved into a traditional Introductory Biology II course. The cohort of first-semester *ICB* students as a whole performed well (83.5%) and equivalent to peer-control students (83.2%; Figure 4B). In addition, a small group of 10 students were examined who had entered the *ICB* curriculum in the second semester after completing a traditional first-semester Introductory Biology I course. The average final grade of the cohort of 10 students was lower, but not significantly so (73.85% vs. 77.48%, $p = 0.4361$), than peers from a full yearlong *ICB* experience (Figure 4C).

DISCUSSION

Assessing Gains in Expertise

We hypothesized the *ICB* curriculum could boost conceptual expertise of students beyond those learning gains obtained with a traditional textbook–driven curriculum (NRC, 2003; AAMC-HHMI, 2009; AAAS, 2011; Waldrop and Miller, 2015). Hence, during the yearlong intervention, both the BCI (Klymkowsky and Garvin-Doxas, 2008) and BCST (Smith *et al.*, 2013) were used to detect any movement of students along the spectrum from novice to expert (Bedard and Chi, 1992; Chi, 2006). The BCI was selected because it assessed student mastery of fundamental concepts in both organismal and cellular topics when confronting strong distractors based on established frequent misconceptions (Klymkowsky and Garvin-Doxas, 2008; Klymkowsky *et al.*, 2010). The cohort of students enrolled in a full year of *ICB* curriculum had a much greater gain on the concept inventory than those in a traditional biology course in all topics tested (overall +43.19 ± 7.02%, $p < 0.01$). The BCST challenged students to categorize biological scenarios to determine whether they were able to see past the superficial and into deeper conceptual linkages (Smith *et al.*, 2013). In this case, the *ICB* students performed equivalently to controls. Only in the details of subcategories did we found any differences. *ICB* students did have higher achievement for the topic of structure and function (77% vs. 28%), but they had lower achievement for evolution compared with the control courses. For both instruments, discrimination occurred at the posttest stage. Given that the BCI detected strong differentiation and the BCST detected little, the overall finding for gains in expertise is likely positive, but perhaps not as glaringly so as the BCI data might suggest.

Assessing Gains in Content Knowledge

We also hypothesized that the *ICB* curriculum, which is not as explicitly content focused as that driven by traditional, more encyclopedic textbooks, might negatively impact short-term gains in content knowledge. The performance of *ICB* students on the MCAT content posttest indicated gains equal to those of the top quintile of the range achieved by previous semesters since the year 2000. The performance of *ICB* students on each topic tested also indicated gains were within norms. In the *ICB* curriculum, perhaps the structured nature to the course, and focus on pursuing inquiry in lecture enabled students to master a greater percentage of content provided (Freeman *et al.*, 2011). While more encyclopedic textbooks may include a greater percentage of content per page (Rissing, 2013), students using them did not appear to master significantly more content as a result. The data suggest that a full year of *ICB* textbook–driven curriculum led to content knowledge gains equivalent to those seen in the traditional curriculum historically.

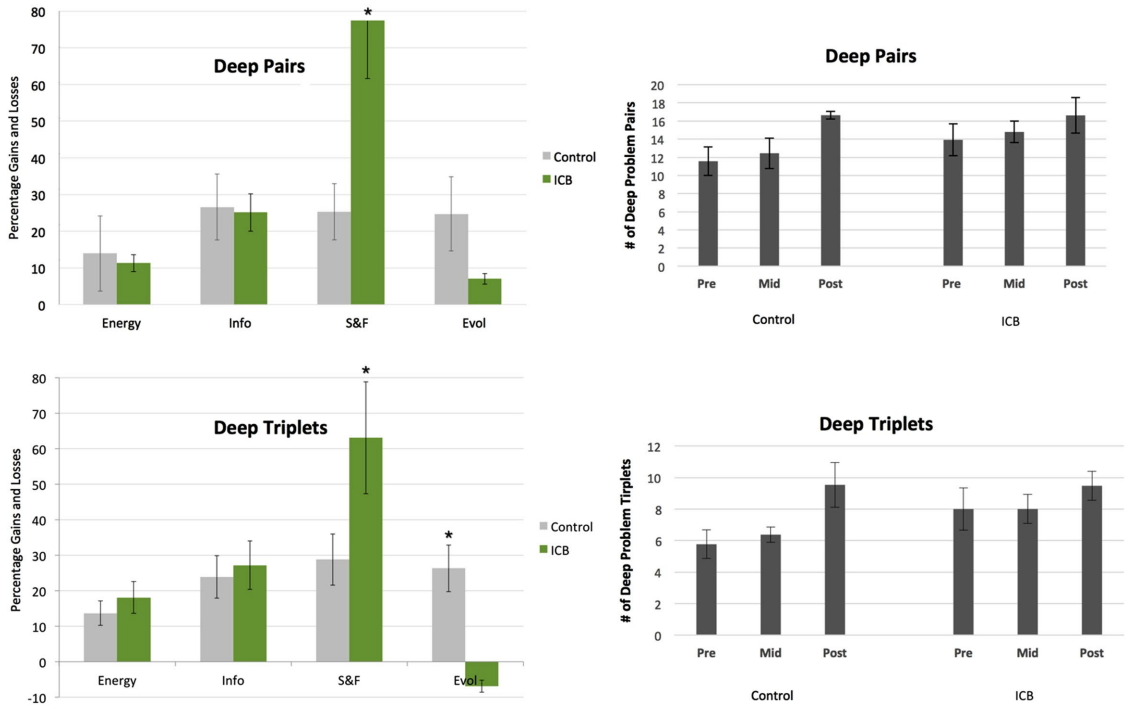

FIGURE 2. BCST performance of *ICB* students and controls. Left, Student normalized performance on posttest for each topic tested. Left to right categories are pathways and transformations of energy and matter, storage and passage of information, structure and function, evolution by natural selection. Right, Student performance at pre-, mid-, and posttest stages for deep pairs and triplets count. The maximum number of deep pairs is 24, and 16 for deep triplets. *ICB* students ($n = 21$) and controls ($n = 101$). Error bars = SEM; *$p < 0.05$.

Detecting Long-Term Impact

We hypothesized the *ICB* curriculum could boost longitudinal performance. Thus, longitudinal tracking was used to look for later success (Voorhees and Lee, 2009; Creech and Sweeder, 2012). The following year, students enrolled in the *ICB* curriculum were tracked into five upper-level courses. Physiology and biochemistry courses were selected due to the number of *ICB* students enrolled and similar academic content. The physics course was selected as a nonbiology control. In addition to ACT normalization, review of college grade point average returned no significant difference between *ICB* student groups in each upper-level course studied. Within one of the physiology courses, the *p* value was significant, indicating that the *ICB* students earned significantly higher grades than their peers. In addition, in one biochemistry course, a greater mean for the *ICB* cohort approached significance, $p = 0.098$. The three other courses lacked significance and showed a lower average grade for the *ICB* group. Yet each had a single outlier point (physics in particular) that, if culled, led to equity of means. This analysis was able to detect early evidence of positive longitudinal impact of the *ICB* curriculum on students.

But Is It the Textbook?

If this were a drug study, changes in dosage would be critical to determine whether the active agent was the drug itself. As mentioned earlier, for both the concept inventory and card-sorting task, discrimination occurred at the posttest stage. This may be somewhat suggestive that dosage, in the form of a full-year of *ICB* textbook–driven curriculum, played a role for impact to manifest or become detectable. Following this line of thought, as an internal control, we first examined computer logs to determine whether use of (or dosage of) textbook correlated positively with exam scores. *ICB* students identified in the bottom quartile for use of the online textbook, scored significantly lower on course exams than those in the top quartile (exam 1: 76% vs. 87%; exam 2: 67%, vs. 79%; exam 3: 62% vs. 71%; $p < 0.05$). Conversely, those who bought the textbook, read it regularly, and provided thoughtful answers to "Integrating Questions" embedded in each textbook reading, scored significantly higher (~10%) on each course exam. As a result of detecting this correlation between dosage of textbook and course exam performance, as a second internal control, we performed a post hoc analysis of longitudinal tracking data. Computer logs were used to identify a small subcohort (9%) of students who never purchased or opened the *ICB* textbook, or answered "Integrating Questions," which served as daily online homework. Empty symbols were used to visualize these "low-dose" students in Figure 4A. This is not empirical data, and we have no comparative data regarding textbook use by peer-control students in their introductory biology courses; thus, it was not used in any of the statistical testing. One would hope that robust use of any textbook would help students in introductory STEM courses. Yet visualizing the trend seemed useful; it is interesting and also somewhat supportive of a potential dosage effect of the *ICB* "drug."

What Makes the Textbook So Special?

The *ICB* textbook is a vast departure from even the most pedagogically progressive traditional textbooks published today. Each chapter reading in the *ICB* textbook focuses on a discussion

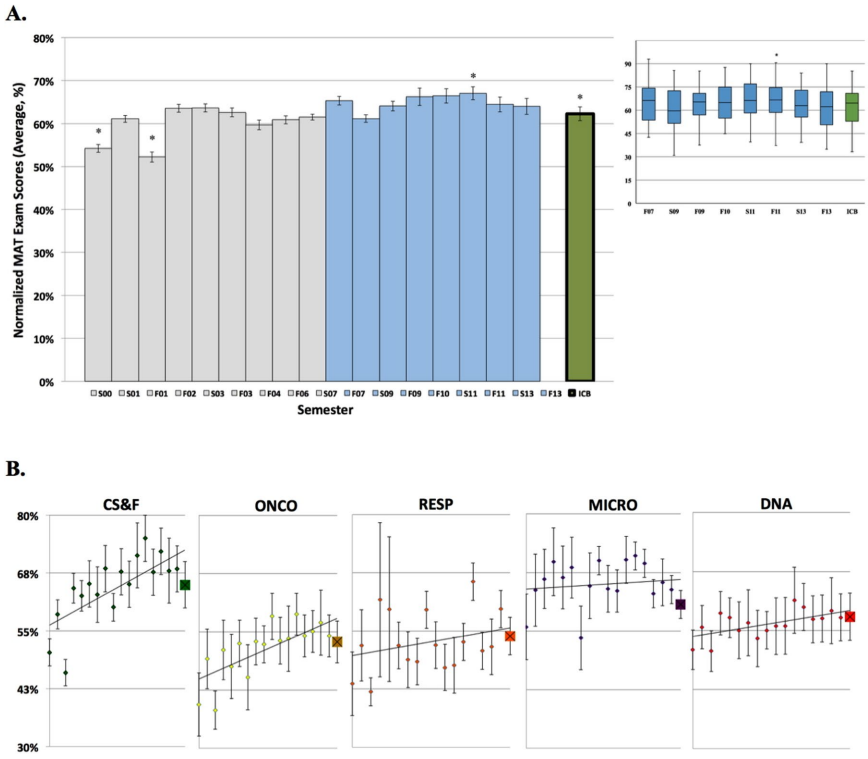

FIGURE 3. ***ICB*** **student cohort performance on MCAT instrument compared with historical scores between 2000 and 2013. (A) Average normalized MCAT performance each semester compared with *ICB* cohort (far right). Error bars = SEM. Inset, Box plot distribution of recent scores: medians and quartiles (box) and range (whiskers). (B) *ICB* performance in each subcategory compared with that of historical controls (left to right: cell structure and function, oncogenes, respiration, microbiology, DNA). Error bars = SEM. Statistical significance is only present when comparing *ICB* with cohorts from years 2000–2001 and 2011. *ICB* students ($n = 77$) and controls ($n = 417$). *$p < 0.05$.**

and dissection of published figures and tables from the primary literature. It then connects that data, from both contemporary and historical researchers, to the fundamental themes and topics of biology. It is unlike the appealingly photogenic and encyclopedic Freeman or Reece biology textbooks with which most veteran instructors are familiar. Rather, it is more akin to a set of published research papers disguised as a textbook to gain conventional acceptance as a source of authority from instructors and students alike. While one might think the departure from a traditional textbook would be uncomfortable for science faculty, in truth, the transition is as familiar as walking from the lecture hall to the professional laboratory. Lectures with undergraduates evolve to be more like journal clubs and have been referred to as "inquiry-in-lecture." In fact, even the conversations after lecture with *ICB* students suddenly sound more like those you have with staff in your research lab. To the authors, interactions with students in the classroom became less frequently like grade school and more like graduate school.

Do No Harm?

When changing curricula, thoughtful skeptics frequently raise Hippocratic concerns that the new approaches not harm the students involved. Because our yearlong *ICB* curriculum did not require students to take both semesters, we were curious, would students who depart (drop out) early, or join (enroll) late, be negatively impacted? According to the p values obtained by the two-tailed t tests, there is no statistical evidence to support a concern that taking only a single semester of *ICB* curriculum may be detrimental to students. Students who did not participate in its entirety, in the full-year dosage of the *ICB* course, were still successful at staying within the boundaries of their peers' performance.

Strengths and Limitations

- One important aspect worth noting is what a course values and achieves (in terms of learning and student effort) is reflected largely through the questions asked on exams. Any comparison of courses must clearly reflect what is to be learned and how that learning is monitored. As a result, this research group is currently completing a project that explicitly characterizes what exactly students are expected to learn and how that is determined via a comparative analysis of exam questions using the 3D-LAP strategy (Laverty *et al.*, 2016).
- BCI: Concept inventories are carefully and professionally developed over years and have to be valid and reliable, but they are also considered particularly difficult for students to do well on and make gains on (Smith and Tanner, 2010). This is a result of each foil being a well-known misconception. Hence, all the wrong answers are particularly attractive to students, that is, they are very strong distractors. For example, in Figure 1A, you will notice that there is no learning gain whatsoever for control students after a full academic year of introductory biology taught by an instructor using reformed engaging techniques. In comparison, the *ICB* students ($n = 76$) did significantly better on the posttest compared with controls ($n = 98$), and in Figure 1B, they did so in every single category tested. The control group that took the concept inventory was the most similar to the experimental group of those present for the different instruments and findings reported in this study. Control students were in the same size class and met in the same room for the same duration on the same days of the week. The student population was homogeneous for both courses; they were students in the same major who all lived in a residential college. While the instructor was not the same, both instructors had the same learning goals, scored similarly on RTOP teaching observations (levels II–IV), and regularly used reformed pedagogies in the same lecture classroom. The same BCI test was taken three times throughout the full academic year, and it therefore does have limitations (also true of the BCST). It

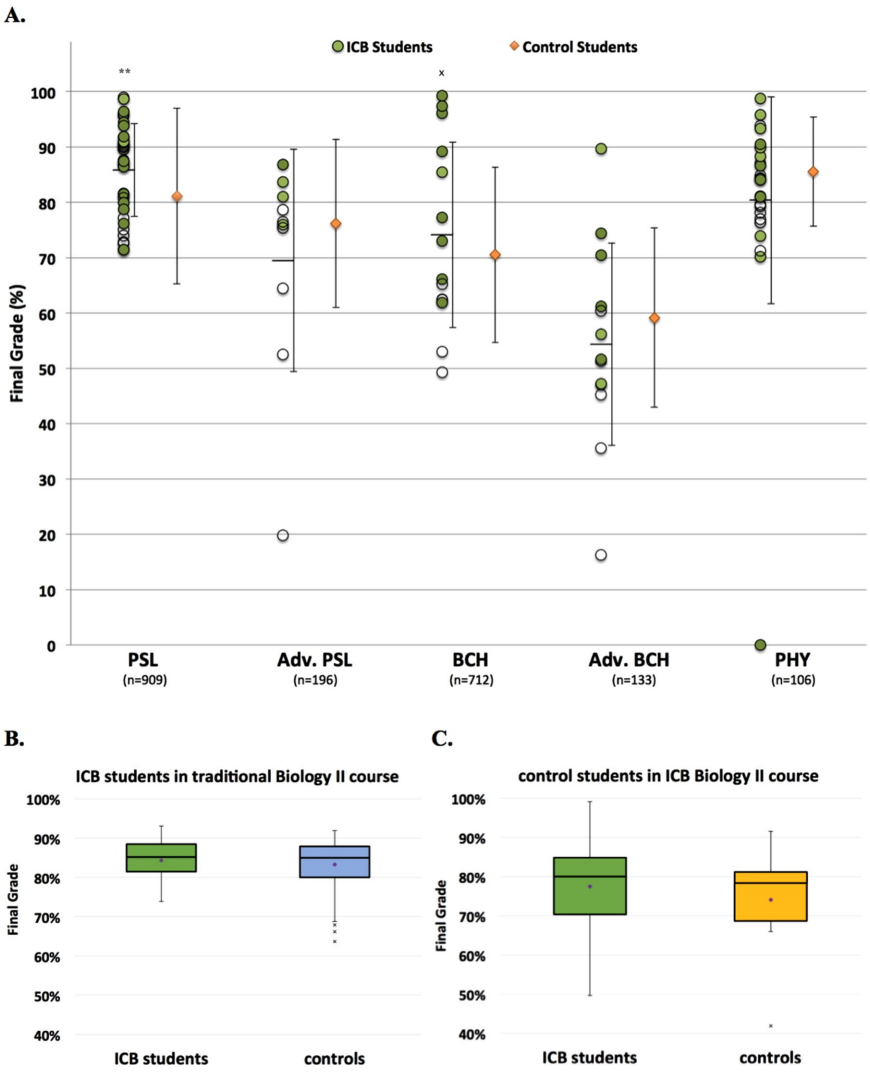

FIGURE 4. Longitudinal tracking of *ICB* student performance vs. peers. (A) Performance the following year in five upper-level STEM courses. Each *ICB* student is represented as an individual point (circles) to the left of control student cohorts (diamonds); "low-dose" students are empty symbols. Error bars = SD. *ICB* students in PSL ($n = 34$), Adv. PSL ($n = 11$), BCH ($n = 26$), Adv. BCH ($n = 13$), and PHY ($n = 24$). **$p < 0.01$; $^{x}p = 0.098$. (B) Box plots of grade performance for first-semester *ICB* students ($n = 25$, left, *ICB* students) vs. peers ($n = 74$, right, controls) in second-semester traditional Introductory Biology II course. (C) Box plots of grade performance for first-semester traditional students ($n = 10$, right, controls) vs. *ICB* peers ($n = 68$, left, *ICB* students) in second-semester *ICB* Introductory Biology II course. Box plot distribution: medians and quartiles (box) and range (whiskers); mean = diamond symbol.

is possible students remembered the questions and became better at answering them by the time they took the posttest. Yet impact at midtest was not seen, and subtracting controls should eliminate common inflation effects.

- BCST: The BCST data set indicated no overall significance in performance differences between *ICB* and control students in both the deep pair and triplet posttest. The finer-grained analysis of performance on individual topics of the card-sorting task revealed that the most significant gain in both pair and triplets was in the category of structure and function, yet there was a loss for evolution by natural selection, which was found statistically significant not in pairs, but in triplets. The greatest limitation to the BCST data is that the trial number is quite low for this assessment ($n = 21$) and the coverage and RTOP in control courses varied widely.
- MAT: This content posttest has an unusual strength in that it has been used for more than a decade and thus is uniquely positioned to compare the impact of a new intervention within the history of the same course. In addition, while the controls are all historical classes with different students, from Fall 2007 to the present ($n = 417$; including the *ICB* semesters, $n = 77$), the instructor remained the same. Hence, in the recent decade, the historical control students experienced the same instructor, classroom, lecture hours, and a homogeneous student population with same major and who all lived in a residential college that was then normalized with ACT scores. On the other hand, the instrument, nicknamed the MAT, was not built with standards of validity and reliability to approach the level of a professional instrument like the BCI or BCST.

Final Thoughts

Perhaps we should not be surprised that a constructivist approach based on science practices, grounded in learning theory, and recommended by experts would improve learning (NRC, 2003; AAMC-HHMI, 2009; AAAS, 2011; Waldrop and Miller, 2015). Rather, the challenge may in fact be detecting and gathering empirical evidence that impact has occurred. It may take a while for the effects of courses that focus on a few core concepts rather than covering a list of topics to show up. In addition, it may be more difficult to measure (using present methods) the benefits of courses that focus on scientific practices, or students doing the same things that scientists do (working with data, argumentation, proposing investigations, modeling, etc.).

While the signal-to-noise ratio was greatest with the concept inventory, which provided great support for enhanced learning by *ICB* students, longitudinal tracking data moderately support positive gains for *ICB* students compared with controls. The MCAT data suggest that there is no sacrifice in learning of "content," and tracking data also indicate that a single semester of *ICB* textbook–driven curriculum did no harm. The data triangulation of three instruments, followed by early longitudinal tracking data, supports that the *ICB* textbook's focus on inquiry, as prescribed by *Vision and Change* (AAAS, 2011), increased learning. Given this cohort of students experienced the first

offering of the *ICB* curriculum at our university, one might predict that going forward, as instructors gain more comfort and experience with the *ICB* textbook and develop more effective approaches to engage students in scientific practices (i.e., inquiry in lecture), the performance of *ICB* students on these instruments may improve further.

ACKNOWLEDGMENTS

We thank Drs. Kendra Cheruvelil, James Smith, Peter White, Gerald Urquhart, and Cheryl Murphy for helpful discussions about teaching and learning and assistance during this study. We also thank the many teaching assistants who promoted and supported this intervention in curriculum and made a fantastic environment for learning in the classroom. Our cystic fibrosis and STEM learning laboratories are supported by the Cystic Fibrosis Foundation (CFF) and Pennsylvania Cystic Fibrosis Inc. (PACFI) grants for our physiology cystic fibrosis research program and by grants from the National Science Foundation for our education research program.

REFERENCES

Adams WK, Wieman C (2011). Development and validation of instruments to measure learning of expert-like thinking. Int J Sci Educ 33, 1289.

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

Anderson DL, Fisher KM, Norman GJ (2002). Development and evaluation of the Conceptual Inventory of Natural Selection. J Res Sci Teach 39, 952–978.

Association of American Medical Colleges and the Howard Hughes Medical Institute (2009). Report of Scientific Foundations for Future Physicians Committee, Washington, DC: Association of American Medical Colleges.

Barsoum M, Sellers PJ, Campbell AM, Heyer LJ, Paradise CJ (2013). Implementing recommendations for introductory biology by writing a new textbook. CBE Life Sci Educ 12, 106–116.

Bedard J, Chi MTH (1992). Expertise. Curr Dir Psychol Sci 1, 135–139.

Campbell AM, Heyer LJ, Paradise CJ (2014). Integrating Concepts in Biology, 1st ed., Davie, FL: Trunity. www.trunity.com/products/digital-textbooks/integrating-concepts-in-biology (accessed 16 March 2017).

Chi MT, Feltovich PJ, Glaser R (1981). Categorization and representation of physics problems of experts and novices. Cogn Sci 5, 121–152.

Chi MTH (2006). Methods to assess the representations of experts' and novices' knowledge. In: Cambridge Handbook of Expertise and Expert Performance, ed. KA Ericsson, N Charness, P Feltovich, and R Hoffman, New York: Cambridge University Press, 167–184.

Creech LR, Sweeder RD (2012). Analysis of student performance in large-enrollment life science courses. CBE Life Sci Educ 11, 386–391.

Derting LR, Sweeder RD (2012). Learner-centered inquiry in undergraduate biology: positive relationships with long-term student achievement. CBE Life Sci Educ 9, 462–472.

Donnon T, Paolucci E, Violato C (2007). The predictive validity of the MCAT for medical school performance and medical board licensing examinations: a meta-analysis of the published research. Acad Med 82, 100–106.

Dreyfus HL, Dreyfus SE (2005). Peripheral vision expertise in real world contexts. Organ Stud 26, 779–792.

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.

Freeman S (2010). Biological Science, 4th ed., San Francisco, CA: Pearson.

Freeman S, Haak D, Wenderoth MP (2011). Increased course structure improves performance in introductory biology. CBE Life Sci Educ 10, 175–186.

Freeman S, Quillin K, Allison L (2014). Biological Science, 5th ed., San Francisco, CA: Pearson.

Garvin-Doxas K, Klymkowsky MW (2008). Understanding randomness and its impact on student learning: lessons learned from building the Biology Concept Inventory (BCI). CBE Life Sci Educ 7, 227–233.

Hake RR (1998). Interactive-engagement versus traditional methods: a six-thousand-student survey of mechanics test data for introductory physics courses. Am J Phys 66, 64–74.

Helldén G (2005). Exploring understandings and responses to science: a program of longitudinal studies. Res Sci Educ 35, 99–122.

Hoskinson A-M, Maher JM, Bekkering C, Ebert-May D (2017). A problem-sorting task detects changes in undergraduate biological expertise over a single semester. CBE Life Sci Educ 16, ar21.

Jeffreys MR (2007). Tracking students through program entry, progression, graduation, and licensure: assessing undergraduate nursing student retention and success. Nurse Educ Today 27, 406–419.

Klymkowsky MW, Garvin-Doxas K (2008). Recognizing student misconceptions through Ed's tools and the Biology Concept Inventory. PLoS Biol 6, e3.

Klymkowsky MW, Underwood SM, Garvin-Doxas RK (2010). Biological Concepts Instrument (BCI): a diagnostic tool for revealing student thinking. arXiv 1012.4501.

Larkin JH, Reif F (1979). Understanding and teaching problem-solving in physics. Eur J Sci Educ 1, 191–203.

Laverty JT, Underwood SM, Matz RL, Posey LA, Carmel JH, Caballero MD (2016). Characterizing college science assessments: the three-dimensional learning assessment protocol. PLoS One 11, e0162333.

Luckie DB, Aubry JR, Marengo BJ, Rivkin AM, Foos LA, Maleszewski JJ (2012). Less teaching, more learning: 10-yr study supports increasing student learning through less coverage and more inquiry. Adv Physiol Educ 36, 325–335.

Luckie DB, Maleszewski JJ, Loznak SD, Krha M (2004). Infusion of collaborative inquiry throughout a biology curriculum increases student learning: a four-year study of "Teams and Streams." Adv Physiol Educ 28, 199–209.

Luckie DB, Rivkin AM, Aubry JR, Marengo BJ, Creech LR, Sweeder RD (2013). Verbal final exam in introductory biology yields gains in student content knowledge and longitudinal performance. CBE Life Sci Educ 12, 515–529.

Newell A, Simon HA (1972). Human Problem Solving, Englewood Cliffs, NJ: Prentice-Hall.

Newman DL, Snyder CW, Fisk JN, Wright LK (2016). Development of the Central Dogma Concept Inventory (CDCI) assessment tool. CBE Life Sci Educ 15, ar9.

National Research Council (NRC) (2003). BIO 2010: Transforming Undergraduate Education for Future Research Biologists, Washington, DC.

NRC (2012). A Framework for K–12 Science Education: Practices, Crosscutting Concepts, and Core Ideas, Washington, DC: National Academies Press.

NRC (2014). Developing Assessments for the Next Generation Science Standards, Washington, DC: National Academies Press.

Next Generation Science Standards Lead States (2013). Next Generation Science Standards: For States, By States, Washington, DC: National Academies Press.

Oakes J (1992). Can tracking research inform practice? Technical, normative, and political considerations. Educ Res 21, 12–21.

Prestwich KN, Sheehy AM (2015). Integrating Concepts in Biology: a model for more effective ways to introduce students to biology. CBE Life Sci Educ 14, fe3.

Raven P, Johnson G, Mason K, Losos J, Singer S (2011). Biology, 9th ed., New York: McGraw-Hill.

Reece JB, Urry LA, Cain ML, Wasserman SA, Minorsky PV, Jackson RB (2013). Campbell Biology, 10th ed., San Francisco, CA: Pearson.

Rissing SW (2013). Correlation between MCAT biology content specifications and topic scope and sequence of general education college biology textbooks. CBE Life Sci Educ 12, 429–440.

Sawada D, Piburn M, 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 Sci Math 102, 245–253.

Smith JI, Combs ED, Nagami PH, Alto VM, Goh HG, Gourdet MA, Hough CM, Nickell AE, Peer AG, Coley JD*, et al.* (2013). Development of the Biology Card Sorting Task to measure conceptual expertise in biology. CBE Life Sci Educ 12, 628–644.

Smith JI, Tanner K (2010). The problem of revealing how students think: concept inventories and beyond. CBE Life Sci Educ 9, 1–5.

Smith MK, Wood WB, Knight JK (2008). The Genetics Concept Assessment: a new concept inventory for gauging student understanding of genetics. CBE Life Sci Educ 7, 422–430.

Smith MU, Good R (1984). Problem solving and classical genetics: successful versus unsuccessful performance. J Res Sci 21, 895–912.

Voorhees R, Lee J (2009). Basics of Longitudinal Cohort Analysis: Principles and Practices of Student Success. Achieving the Dream: Community Colleges Count. Lumina Foundation for Education. files.eric.ed.gov/fulltext/ED532373.pdf (accessed 16 March 2017).

Wagner JD, Campbell AM, Sly BJ, Paradise CJ (2015). An active textbook converts "vision and tweak" to *Vision and Change*. CourseSource. www.coursesource.org/courses/an-active-textbook-converts-vision-and-tweak-to-vision-and-change#tabs-0-content=4 (accessed 16 March 2017).

Wai J, Lubinski D, Benbow C, Steiger J (2010). Accomplishment in science, technology, engineering, and mathematics (STEM) and its relation to STEM educational dose: a 25-year longitudinal study. J Educ Psychol 102, 860–871.

Waldrop LD, Miller LA (2015). Introduction to the symposium "Leading Students and Faculty to Quantitative Biology through Active Learning." Integr Comp Biol 55, 898–900.

White RT, Arzi HJ (2005). Longitudinal studies: designs, validity, practicality, and value. Res Sci Educ 35, 137–149.

Zheng AY, Lawhorn JK, Lumley T, Freeman S (2008). Application of Bloom's taxonomy debunks the "MCAT myth." Science 319, 414–415.