A Problem-Sorting Task Detects Changes in Undergraduate Biological Expertise over a Single Semester

Anne-Marie Hoskinson,[†]* Jessica Middlemis Maher,[‡] Cody Bekkering,[§] and Diane Ebert-May[§]

[†]Department of Biology and Microbiology, South Dakota State University, Brookings, SD 57007; [‡]Delta Program in Research, Teaching, and Learning, University of Wisconsin–Madison, Madison, WI 53706; [§]Department of Plant Biology, Michigan State University, East Lansing, MI 48824

ABSTRACT

Calls for undergraduate biology reform share similar goals: to produce people who can organize, use, connect, and communicate about biological knowledge. Achieving these goals requires students to gain disciplinary expertise. Experts organize, access, and apply disciplinary knowledge differently than novices, and expertise is measurable. By asking introductory biology students to sort biological problems, we investigated whether they changed how they organized and linked biological ideas over one semester of introductory biology. We administered the Biology Card Sorting Task to 751 students enrolled in their first or second introductory biology course focusing on either cellular–molecular or organismal–population topics, under structured or unstructured sorting conditions. Students used a combination of superficial, deep, and yet-uncharacterized ways of organizing and connecting biological knowledge. In some cases, this translated to more expert-like ways of organizing knowledge over a single semester, best predicted by whether students were enrolled in their first or second semester of biology and by the sorting condition completed. In addition to illuminating differences between novices and experts, our results show that card sorting is a robust way of detecting changes in novices' biological expertise—even in heterogeneous populations of novice biology students over the time span of a single semester.

INTRODUCTION

The past 30 years have seen many discussions, publications, and policy positions advanced around the question "What should be the purposes of undergraduate science education?" The American Association for the Advancement of Science's *Science for All Americans* (AAAS, 1991; Rutherford and Ahlgren, 1991) used "habits of mind" to describe the suite of scientific, mathematical, and logical thinking skills that young people should adopt during their school years. The Committee on Undergraduate Science Education (National Research Council [NRC], 1999), justified then-new National Science Education standards, in part, by explaining,

> "Citizens need scientific information and ways of thinking in order to make informed decisions, [and] business and industry need … workers with the ability to learn, reason, think creatively, make decisions, and solve problems." (NRC, 1999, p. 2)

Brewer and Smith reiterated and reframed these and other goals in *Vision and Change in Undergraduate Biology Education: A Call to Action* when they wrote,

> "Biology in the 21st century (NRC, 2009) requires that undergraduates learn how to integrate concepts across levels of organization and complexity and to synthesize and analyze information that connects conceptual domains." (AAAS, 2011)

Erin L. Dolan, *Monitoring Editor*

Submitted May 24, 2016; Revised December 6, 2016; Accepted January 17, 2017

CBE Life Sci Educ June 1, 2017 16:ar21

DOI:10.1187/cbe.16-05-0175

*Address correspondence to: Anne-Marie Hoskinson (annemarie.hoskinson@sdstate.edu).

These are remarkably common aspirations for undergraduate science education. Indeed, *Vision and Change* identified a set of five "core concepts" (AAAS, 2011, pp. 11–13) and six "core competencies" (AAAS, 2011, p. 17) as a framework for guiding undergraduate biology education and student learning. These performance goals focus on developing individuals who can organize, use, make connections among, and communicate about biological knowledge and information. One way to characterize these and other works is to say they envision students approaching *disciplinary expertise* (Chi *et al.*, 1981; Glaser and Chi, 1988; Shanteau, 1992). What, then, is an expert? How do instructors and students produce learning with deep understanding and connectivity, and beyond merely fact acquisition? These are open questions in biology education research. In this paper, we present an argument for viewing biological learning through a lens of expertise, using empirical data collected to support our claims.

Cognitive scientists have a rich history of studying expertise among experts in many domains, from expert performers in sport and music, to cognitive experts in medicine, law, and air traffic control. Despite domains that can make very different demands on performers and that require demonstrating different skills or knowledge, experts display a remarkably consistent set of characteristics. Bransford and colleagues (NRC, 2000, p. 31) summarize six properties; for the purposes of the present investigation, we concentrate on three. First, experts extract and use *meaningful patterns* in information and among individual components of systems that novices do not. Second, expert knowledge is situated in the context of the expert's environment; his or her past history, learning, and knowledge; and the applicability or usefulness of that knowledge to the present situation. This is what Bransford and colleagues termed "*conditionalized* on a set of circumstances." Finally, experts both *know a great deal more* than novices, and they use *deep principles to organize* and access what they know. In terms of our biology students and our undergraduate classrooms, these properties closely reflect visions of people who: possess *knowledge* of "scientific information and ways of thinking" (NRC, 1999); can use their knowledge *conditionally* to "think creatively, make decisions, and solve problems"; and "synthesize and analyze information that connects conceptual domains" using *deep principles* (AAAS, 2011, p. ix). Importantly, expertise is domain specific—that is, expertise in biology cannot be built or transferred from expertise in other disciplines.

Early empirical investigations of cognitive expertise focused on master chess players (deGroot, 1965; Chase and Simon, 1973). DeGroot demonstrated that expert chess players could reconstruct board configurations accurately, even when they saw the configuration only for a few seconds, whereas novices were much less successful or accurate at this task. Chase and Simon (1973) built on this work by showing that experts perform "chunking"—that is, aggregating the identities and positions of up to 32 pieces on a 64-square playing board—into many fewer, internally cohesive blocks of information. These chunks account in part for experts' abilities to notice meaningful patterns. Under chunking theory, expert chunks in memory are more numerous, larger, and better indexed than those of novices. Experts use heuristics, or rules of thumb, that increase the efficiency of searches for the most-relevant chunks. Experts create these heuristics over many experiences and multiple exposures to patterns until they can recognize many patterns much more rapidly than novices. This feature of chunking theory helps to explain experts' conditionalized knowledge: their ability to rapidly and efficiently access specific knowledge in a given context.

As chunking theory was tested and improved, researchers also demonstrated that experts have multiple, dynamic frameworks that can also give rise to higher-level concepts—in the present study, core concepts of biology (Gobet and Simon, 1996). This helped to explain the third characteristic of experts, who organize their knowledge based on the deep concepts or principles of their disciplines. This improvement, currently dubbed *template theory*, includes several propositions for processes that could help to explain how people develop expertise generally, and specifically in our biology classrooms. The first proposition is the availability of an expert-like organizational template for novices. For novices, access to an expert-like template may help accomplish two important goals. First, it provides an incomplete structure to use, test, and adapt to the information and stimuli of the learner's environments; and second, it can constrain what the learner pays attention to, thereby helping develop the conditionalized quality of expert knowledge. The next proposition of chunking/template theory is that there may be a dosage effect. This is termed *deliberate practice* in some literature. The dosage effect posits that time investment is a necessary cost but not a sufficient requirement for developing expertise. Cognitive scientists have carefully studied and quantified very short times required for perception and memory access, and the very long times required to develop expertise (usually in tens of thousands of hours or in years; Ericsson *et al.*, 1993; Hambrick *et al.*, 2014). There is a large gap in studying whether there is a dosage effect for putative novices advancing toward expertise, cognizant that some individuals may never achieve true mastery or expertise (de Bruin *et al.*, 2014), which is characteristic of many populations of introductory biology students. Finally, chunking and template theories suggest that the order of concepts in learning might be important in developing expertise, as perceptual skills are developed from concrete concepts, then honed with abstract concepts. For example, Gobet (2005) noted from studies of chess players that learning from concrete concepts to abstract concepts best facilitated their advancement toward expertise. Other workers studying concepts in education have noted that there are different types of concepts (Lawson *et al.*, 2000; Medin *et al.*, 2000) and that people engage with different types of ideas and concepts differently (Solomon *et al.*, 1999; Lawson *et al.*, 2000). Among biology instructors, there exist varying beliefs but a paucity of research (e.g., Michael, 2007) about which concepts and topics are easier or harder for students to learn.

Our approach to studying expertise in the present study draws on a technique currently dubbed "card sorting." Card sorting has its origins in early investigations of knowledge acquisition and concept formation. In a card sort, subjects are presented with cards displaying objects, shapes, words, or ideas, and they are then asked to sort or group similar cards together into meaningful groups. Investigators might be interested either in subjects' processes of forming groups, on the products of the sorting task, or both. Depending on the experimental conditions, subjects might be given group labels a priori, termed a *closed* sort (Fincher and Tenenberg, 2005), in order to probe subjects' abilities to recognize and use a given organizational framework. Alternatively, subjects might be asked both to sort

	Energy and Matter (E & M)	Information Storage and Transfer (Info)	Structure and Function (S & F)	Evolution by Natural Selection (Evol)
Plant	D	J	I	K
Insect	H	B	M	F
Human	L	O	P	N
Micro-organism	A	E	G	C

FIGURE 1. The 16 problems in the BCST, each having a deep-concept feature (*Vision and Change* core concepts; columns) and a superficial feature (organisms; rows).

the cards and to assign a label to each of the groups they create, termed an *open* sort (Fincher and Tenenberg, 2005). Here, the purpose is to discover how subjects, novices and experts alike, differ in how they link and organize ideas.

Chi and colleagues used a physics problem-sorting task to characterize differences between novices and experts in physics (Chi *et al.*, 1981). They investigated what categories people used to sort physics textbook problems and how the categories differed among putative novices (undergraduate physics students) and experts (advanced PhD students). They found that novices tended to categorize problems based on *superficial* features: objects, physics terms, or problem features (Chi *et al.*, 1981; Figure 1). Experts tended to categorize problems by the deeper physical *principles or concepts* of the problem, such as conservation of energy and Newton's second law (Chi *et al.*, 1981; Figure 2). Subsequent work on students' sorting of mathematics problems (Schoenfeld and Herrmann, 1982) and computer science problems (McCauley *et al.*, 2005; Sanders *et al.*, 2005) support this key distinction between novice and expert frameworks.

Smith and her colleagues used this distinction between novice and expert frameworks to design the Biology Card Sorting Task (BCST; Smith *et al.*, 2013). Their purpose in designing and testing the BCST was to quantitatively measure biology concept expertise in multiple ways using a hypothesis-driven design. Their hypothesized deep concepts were taken from *Vision and Change* (AAAS, 2011) and represented four of the five proposed deep concepts: pathways and transformations of energy and matter (herein, "E&M"), storage and passage of information ("Info"), relationships between structure and function ("S&F"); and Evolution ("Evol"). (They integrated the fifth *Vision and Change* core concept, systems, within their four hypothesized deep categories.) Hypothesized superficial features were represented by groups of organisms: humans, microorganisms, insects, and plants. Smith and colleagues selected problems from introductory biology texts so that each problem included one and only one superficial feature and one and only one deep-concept feature, and they also chose problems that were not "too difficult," consistent with cautions by Wolf and colleagues (2012). To explore both the products people generated and their processes for generation, the researchers used both open and closed sorting. In the closed-sort condition, which they termed *framed* sorting, the four deep categories from *Vision and Change* form the framing. Smith and colleagues also used an open-sort condition, in which subjects both sorted the problems and labeled the groups into which they sorted the problems, termed *unframed* sorting. They iterated several cycles of problem selection, tool development, administration, and think-aloud interviews with two subject pools: putative novices (introductory, non-biology majors in their first college biology course) and putative experts (tenure-track college biology faculty).

Smith and colleagues (2013) extracted several key findings. First, the experts sorted biology problems according to their deep features significantly more frequently than novices, under both the framed (closed) sorting condition, and the unframed (open) sorting condition. Under both framed and unframed sorting conditions, college biology faculty used deep features of the problems to sort them, whereas introductory biology students used a combination of superficial, deep, and unknown (neither superficial nor deep) features to sort their problems. Under the unframed condition, experts were significantly more likely to label their sort categories according to deep biological principles than novices, who were significantly more likely to assign superficial category labels (e.g., organisms). Finally, expert subjects' sorts were significantly closer to the hypothesized deep-feature sort than novice subjects' sorts.

Left open were the questions of what factors helped novices develop expertise, and whether their tool could capture any aspects of that process. The present study had two overarching purposes. First, we aimed to investigate the applicability of scaling up the BCST "in the wild," to college students enrolled in one-semester, large-enrollment, introductory biology courses. Second, we tested whether and how each of three propositions from chunking and template theory could explain advancement toward expertise. Specifically, we addressed the following research questions:

1. Can the BCST discern population-level changes in expertise over a single introductory biology course? Smith and colleagues were optimistic about the broad applicability, but their putative novices and experts were also distinct from one another by at least eight years of education and professional experiences. We aimed to explore whether the BCST detected changes in student expertise over a single semester of introductory biology and in novice populations.

A) Framed Sorting

Evolution by natural selection	Pathways and transformations of energy and matter	Storage and passage of information about how to build living systems	Relationships between structure and function
F, K, N	A, D, I, L	B, E, J, O	C, G, H, M, P

B) Unframed Sorting

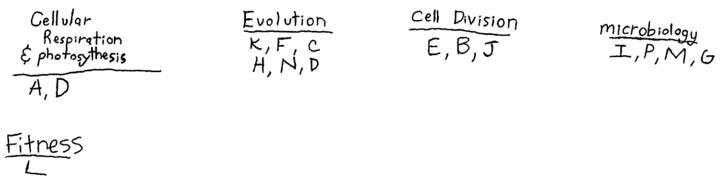

FIGURE 2. Examples of a student's framed sort (A) and a student's unframed sort (B).

2. Which, if any, factors help explain how relative novices develop expertise over a single semester of introductory biology? The present study tests three possible explanatory factors suggested by chunking and template theories: whether availability of an expert framework facilitates advancement toward expertise (the *framework proposition*); whether there is a *dosage effect*; and whether the order of the two introductory *course topics* matters.

METHODS

Subjects

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 at a large midwestern land-grant university. The 751 participants were enrolled in one of 15 different course sections varying in topic and sequence (see *Procedure* section). Many but not all students were biology or other science, technology, engineering, and mathematics majors; in total, 82 different majors were represented in our population. We describe student demographics in Table 1.

The Task

We used a problem-sorting task, developed and tested by Smith and colleagues, composed of 16 biology problems selected from common introductory biology textbooks (Smith *et al.*, 2013). Smith and colleagues modified these problems in ways that eliminated jargon, graphics, and overt cues to core biological concepts. Each problem possessed a single deep-concept feature and a single superficial (Chi *et al.*, 1981) feature. The deep-concept features were adapted from the core biological concepts in *Vision and Change* (Smith *et al.*, 2013) as described in the *Introduction* and summarized here: 1) energy and matter (E&M); 2) information storage and transfer (Info); 3) structure and function (S&F); and 4) evolution (Evol; Figure 1, column headers). Superficial features were groups of organisms: plants, insects, humans, and micro-organisms (Figure 1, row labels). We used Smith and colleagues' 16 vetted problems unmodified, except that we inadvertently transposed two problem identifiers, F and H, before our first administration in Spring 2014. To simplify data collection and analysis, we preserved the problem-ID transposition. To maintain the integrity of the task and limit student access, the problems we used in this study are not included in this publication but can be obtained by contacting the corresponding author for the BCST (Smith *et al.*, 2013). In the following sections, we describe how we used and administered the BCST to respond to our research questions.

Procedure

We administered the BCST in 15 introductory biology courses spanning three semesters using a pre–post design: during the first or second week of the semester and again during the second to last or last week of the semester (Table 2). The pre–post design allowed us to address our first research question of whether the BCST detects changes in student biological expertise over a single semester. We gave each student two sheets of paper with the 16 problems printed and labeled "A" through "P," and one sorting form. We informed students that the purpose of the task was to sort the problems, not to solve them. We gave students unlimited time to read the problems and assign them to categories on their sorting forms, but found that 20 minutes was more than adequate to allow all students to complete the task. Students completed this task individually, without the assistance of textbooks, and without access to the internet or notes. Students also completed a consent form and could opt out of the investigation at any time. This investigation was reviewed, the protocol approved, and the study determined exempt by the university's Institutional Review Board (IRB# x14-026e, File ID: i045259).

We also aimed to investigate three possible explanations for changes in student expertise: the framework proposition, exploring whether there was an effect of sorting under framed (closed) or unframed (open) conditions; the dosage effect, inquiring whether and how first or second courses affected changes in expertise; and the topic proposition, testing whether and how broad course topics affected changes in

TABLE 1. Student demographics by sex, class standing, nonwhite, and international student status with SEMs in parentheses

Course	% Male	% Female	% Freshman	% Sophomore	% Junior	% Senior	% Nonwhite	% International
1	43	57	57	20	17	7	11	2
2	28	72	46	38	13	3	13	4
3	57	43	34	42	20	4	17	5
4	22	78	47	44	9	0	22	0
5	38	62	2	94	5	0	15	0
6	60	40	54	22	13	10	24	6
7	47	53	6	76	6	12	24	6
8	39	61	6	54	35	6	19	3
9	29	71	4	57	30	9	17	0
10	18	82	59	18	10	13	15	0
11	39	61	47	39	7	7	18	5
12	51	49	31	44	13	10	13	0
13	31	69	31	50	13	6	6	0
14	42	58	6	49	39	6	25	0
15	20	80	15	30	55	0	25	10
Mean	38 (3)	62 (3)	30 (6)	45 (5)	19 (4)	6 (1)	18 (1)	3 (1)

TABLE 2. Surface, deep, and unexpected problem pairings for each of the 15 sections of introductory biology included in this study[a]

Course	*N*	Semester	Sort type	Sequence	Topic	% Surface (Pre)	% Deep (Pre)	% Unexpected (Pre)	% Surface (Post)	% Deep (Post)	% Unexpected (Post)
1F	29	1	F	First	Cell-Molec	13 (1)	50 (5)	37 (3)	15 (1)	42 (4)	44 (3)
2F	47	1	F	First	Cell-Molec	14 (1)	42 (4)	43 (3)	13 (1)	52 (4)	36 (3)
3	76	2	F	First	Cell-Molec	17 (1)	41 (3)	43 (2)	13 (1)	41 (3)	46 (3)
4F	22	1	F	First	Org-Pop	15 (2)	43 (6)	42 (5)	10 (2)	59 (7)	30 (5)
5F	39	2	F	First	Org-Pop	11 (1)	57 (4)	32 (3)	11 (1)	56 (4)	33 (3)
6	68	2	F	First	Org-Pop	13 (1)	46 (3)	41 (2)	12 (1)	47 (3)	41 (2)
7F	17	1	F	Second	Org-Pop	13 (1)	46 (5)	41 (4)	10 (2)	64 (6)	27 (4)
8F	44	1	F	Second	Org-Pop	13 (1)	49 (4)	38 (3)	9 (1)	66 (4)	26 (3)
9	70	3	F	Second	Org-Pop	12 (1)	48 (3)	39 (2)	8 (1)	66 (4)	26 (2)
1U	17	1	U	First	Cell-Molec	42 (7)	25 (5)	33 (4)	40 (7)	24 (4)	36 (4)
2U	22	1	U	First	Cell-Molec	37 (8)	36 (6)	27 (4)	26 (5)	40 (5)	34 (3)
10	39	3	U	First	Cell-Molec	41 (5)	30 (4)	29 (2)	30 (5)	39 (4)	31 (3)
11	57	3	U	First	Cell-Molec	37 (4)	34 (3)	29 (2)	36 (4)	36 (3)	29 (2)
4U	10	1	U	First	Org-Pop	26 (6)	38 (6)	36 (4)	13 (3)	50 (7)	36 (4)
5U	27	2	U	First	Org-Pop	39 (7)	30 (5)	31 (4)	44 (6)	29 (5)	26 (3)
12	39	3	U	First	Org-Pop	45 (5)	29 (4)	27 (2)	44 (5)	29 (4)	27 (3)
13	16	2	U	Second	Cell-Molec	30 (6)	45 (7)	25 (4)	12 (2)	58 (4)	30 (2)
8U	25	1	U	Second	Org-Pop	32 (5)	30 (4)	38 (3)	23 (5)	44 (6)	33 (4)
14	67	2	U	Second	Org-Pop	33 (4)	35 (3)	31 (2)	27 (3)	43 (3)	30 (2)
15	20	3	U	Second	Org-Pop	43 (7)	29 (5)	28 (4)	21 (5)	44 (5)	35 (3)

[a]During the first semester of our investigation, we administered both sort types within each course section in a randomized split-block design; these courses are designated with an F or a U after the course number. The count *N* shows valid sorts for that course section (see *Methods*). Under sort type, F indicates the framed sort, and U indicates the unframed sort. Sequence denotes whether the section was a first course or second course in the introductory track. For topic, Cell-Molec indicates courses that focused on cellular and molecular biology, and Org-Pop indicates courses that focused on biology at the organism and population level. Percentage pairings are shown for both the pre and post sorts, with SEMs in parentheses.

expertise. To address the framework proposition, we administered the BCST under one of two sorting conditions: a framed sort (Figure 2A) and an unframed sort (Figure 2B). In the framed sort, we provided students with the four deep-concept categories into which they should sort problems. In the unframed sort, students were not provided with any categories; instead, we asked students to group problems, then create a name for each category. We directed students to create as few as two or as many as 15 categories and to label each category. Cognitively, then, the unframed sort can be quite different from the framed sort, especially for novices (Chi *et al.*, 1981; Gobet and Simon, 1996; Gobet, 1998; Fincher and Tenenberg, 2005). Under both sorting conditions, we directed students to sort all problems based on what they knew about biology and to make sure they sorted each and every problem into one and only one category.

Some students in this study were taking their first introductory biology course, and others were taking their second introductory biology course. This allowed us to test whether there was a dosage effect under our second research question. Introductory biology courses at this university varied in how instructors structured their courses. In some cases, instructors endeavored to align their courses with biological concept (AAAS, 2011; Campbell *et al.*, 2014). In other cases, instructors chose a more topical approach to their course design. Because there was no coordination among most instructors other than partitioning broad topics between the semesters, we define "dosage effect" here in the sense of a larger (longer time) dose, rather than a more complete dose of biology. Hence, our use of "dosage" is consistent with research on deliberate practice, which posits an effect of time on the potential attainment of expertise (Ericsson *et al.*, 1993; Hambrick *et al.*, 2014).

Finally, design of the introductory biology sequence at this university meant that course topics fell into two categories: courses that focused on cellular and molecular biology and courses that focused on organismal and population biology. This division of topics across courses allowed us to test the third proposition, the topic proposition. There were two different introductory biology course tracks at this university: one with cellular–molecular topics in the first semester and organismal–population topics in the second semester, and the other track reversing this sequence. Therefore, course topic and course sequence were independent of one another. Thus, our study design was a within-subjects, full-factorial design, with factors of course topic, sequence, and sort type as the grouping variables ("fixed effects"; see *Analyses*).

During the first semester of our investigation, we administered both sort types within each course section in a randomized split-block design. However, because one aim of our investigation was to characterize individual and population-level changes in expertise, it was important for each student to complete the BCST under the same sorting condition, both pre and post. When we administered the post test, reliance on individual students correctly remembering their pre test condition led to excessive sort-type mismatches between the pre and post, resulting in a high discard rate. Therefore, after the first semester, all students in a course section completed the BCST under the same sorting condition, either framed or unframed (Supplemental Figure S1).

There were some important differences between the purposes of our data collection and those of Smith and colleagues (2013). First, the pre–post design makes our investigation a within-subjects design. Our analyses included only students who consented to participate in the study; who completed both sorts under the same sort condition, pre and post; and who completed valid sorts according to the sort conditions described earlier. Also, we did not ask students to reflect on the sorting conditions, nor did we impose a deadline or keep time. Finally, during the second and third semesters of administering the BCST, and despite not administering the BCST to all possible introductory course sections, we became aware that it was possible for some students to complete the BCST more than once in two different courses as they moved through the introductory curriculum. To account for any possible priming effects (Forster and Davis, 1984), we discarded data from the small number of students who had completed the BCST during a second course. In other words, the data reported here are all from students completing their first and only instantiation of the BCST, where an instantiation is completing both the pre and post conditions in a single semester.

METRICS

To address our research questions and test our propositions, we used two metrics described by Smith and colleagues: problem pairings and edit distance. In addition, because our data set was considerably larger, we were also able to use a third metric: deep-feature problem triplets. We investigated deep-feature problem triplets as potential illuminators of some introductory students' greater expertise ("initiates" in Hoffman, 1996).

Problem Groupings: Pairs and Triplets

When a student placed two problems into the same category, this constituted a problem pair. Three problems placed in the same category formed a problem triplet. When a student put two problems into the same organism-based category, this constituted a superficial pair. Likewise, when a student put two problems into the same deep-concept category, this constituted a deep pair. Students formed unexpected pairs when they put two problems that did not share a common superficial or deep feature into the same category. Unpaired problems were also dubbed unexpected. Students formed deep- and superficial-problem triplets similarly. However, we could no longer apply the term "unexpected" in the same way with triplets, because a triplet that was neither wholly superficial nor wholly deep might contain a combination of superficial, deep, and unexpected pairs that confounded efforts to characterize them. Consider the following examples to illustrate these principles. In the student's problem grouping {CGHMP} shown in Figure 2A, three pairs—GM, GP, and MP—belong to the deep-feature category S&F (Figure 1); two pairs—CG and HM—belong to superficial-feature categories (micro-organisms and insects, respectively; Figure 1). Pairs CH, CM, CP, HP, and GH represent unexpected pairings, since they are paired neither by superficial nor deep features. The same problem grouping contains a deep-feature triplet—GMP (S&F). However, consider the triplet HMP. It contains a superficial pair, HM (insects); a deep pair, MP (S&F); and an unexpected pair, HP, defying categorization as merely "unexpected." We therefore considered only deep triplets in our analyses.

The 16 problems were therefore potentially arranged into 24 deep pairs (six in each of the four deep categories), 24 superficial pairs, and 88 unexpected pairs (including unpaired problems). Similarly, the same 16 problems could form 16 superficial triplets (four in each category) and 16 deep triplets (four in each category). Superficial-feature triplets were rare, especially under the framed sort, so we did not use them in our analyses. Using this scheme of problem pairings, we were able to quantitatively analyze students' problem sorts.

Edit Distance

In addition to problem pairs and triplets, we used another metric, edit distance, to quantify the problem sorts (Deibel *et al.*, 2005). The edit distance to the deep sort (ED_{deep}) is the number of single problems in a student's sort that would need to be moved to categories corresponding to a deep-feature sort. From the sort shown in Figure 2A, to obtain the deep sort, problem C must be moved to the Evol deep-feature category, problem H to E&M, and problem I to S&F. In Figure 2B, problems H and L must be moved to E&M ("cellular respiration and photosynthesis" on that student's response), and problem O to Info ("cell division"). Because both sorts required three problem moves to arrive at the deep-feature sort, $ED_{deep} = 3$. We used Python scripts developed by Smith and colleagues (2013) to calculate each subject's edit distance to a deep sort.

Normalization and Weighting: Deep Pairs, Deep Triplets, and Edit Distance

One aim of this investigation was to compare results on deep-feature pairs, deep triplets, and edit distance to deep sort, each among large-enrollment courses that varied in sort type, sequence, and topic. To do so, we needed to account for different pre test starting values among courses. We therefore normalized gains and losses in our metrics of deep pairs, deep triplets, and edit distance (Hake, 1998). We calculated normalized gains (or losses) toward deep-feature sorting as the ratio of actual gain (or loss) to possible gain (or loss). For example, suppose we aimed to compare changes in the percentage of students assigning expert pair AD to the expert category E&M between two courses. Also suppose that 35% percent of course 1's students assigned pair AD to E&M on their pre test, and 55% on the post. In course 2, suppose that 48% of students assigned pair AD to E&M on the pre, and 68% on the post test. The net gain is 20% in both courses. However, because students in course 2 started closer to the maximum deep-feature sort percentage (48% vs. 35%), their normalized gain is actually greater than that of course 1:

$$\text{Course 1: } 20/(100-35) = 31\%, \quad \text{Course 2: } 20/(100-48) = 38\%$$

We normalized losses using the same principle. For simplicity, we describe these as relative gains or losses in subsequent figures and text. Additionally, courses varied in enrollment size, and we therefore weighted the course means of our metrics—deep pairs, deep triplets, and edit distance—by the number of individuals completing valid sorts (as described earlier) in that course.

ANALYSES

We first performed *k*-means cluster analysis to probe the robustness of the superficial and deep categories. We then used two sets of analyses to address our research questions:

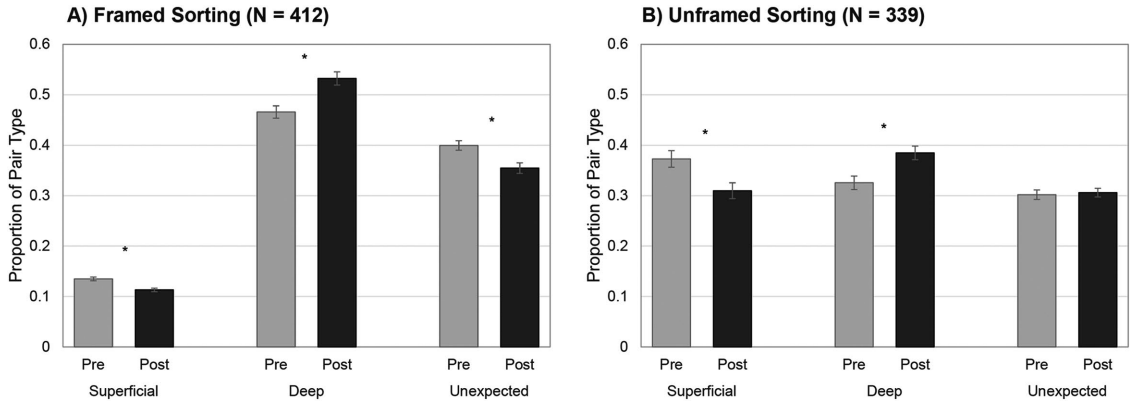

FIGURE 3. Students ($N = 751$) sorted problems significantly differently at the end of a semester of biology (post; black bars) than the beginning (pre; gray bars). Under both framed (A) and unframed (B) sorting conditions, percentage of superficial pairings significantly decreased ($p < 0.001$), while deep pairings significantly increased ($p < 0.001$). Unexpected problem pairings were remarkably similar (30–40%) pre to post and between sort types. Error bars are SEM. Asterisks indicate statistically significant results.

comparative analyses (research question 1) and linear mixed modeling (research question 2). Individual courses were the units of analysis.

Comparative Analyses: Problem Pairs and Triplets

For problem pairs, we calculated the mean percentage of each problem type—superficial, deep, and unexpected—and the total of deep pairs to a maximum of 24. We did this for each sort type (framed and unframed) and for each sort event (pre and post). For problem triplets, we calculated total deep triplets to a maximum of 16. We used two-tailed Student's *t* tests to compare means and Levene's test for equality of variances to select the proper results.

Linear Mixed Modeling: Grouping Variables as Fixed Effects

We used grouping variables describing sort type (framed or unframed), course sequence (first or second), and course topic (cellular–molecular or organismal–population). Each of these three binary variables was potentially a fixed effect, with course section in which the student was enrolled as a random effect in a linear mixed model. Each fixed effect corresponded to a proposition of our second research question: framework proposition (sort type), dosage effect (course sequence), and topic proposition (course topic). We performed linear mixed modeling on normalized change in edit distance to deep sort, normalized gains in deep pairs, and normalized gains in deep triplets to explore each fixed effect—both alone, and in interaction with other fixed effects. In the case of deep pairs and triplets, if one of the grouping variables was a significant factor in LMM, we then performed analysis of variance to determine effects on individual pairs and triplets. Additionally, we calculated effect sizes for all statistically significant or marginally significant results to assess the magnitude of the effect (Maher *et al.*, 2013).

RESULTS

Superficial and Deep Templates/Frameworks

Our results from *k*-means cluster analysis validated superficial and deep categories. Cluster 1 corresponded to deep-feature sorting and contained nine diagnostic deep-problem pairs. Cluster 2, corresponding to superficial sorting, contained 21 superficial-problem pairs (Supplemental Table S1).

Research Question 1: Comparative Analyses of Problem Groupings

In our analysis of card pairings, we observed significant increases in deep pairings and significant decreases in superficial pairings from pre to post in both sorting conditions (Figure 3 and Table 2). Despite being provided the deep-feature category names on the framed sorting task, superficial pairs comprised 13% of total pairs on the pre test, and 11% on the post test (Figure 3A and Table 2). Also, superficial pairings were more common, and deep pairings less common, under the unframed than the framed sorting condition, both pre and post (Figure 3B). Students also made significant gains over a single semester in the number of deep pairs and deep triplets they created under both sort types (Table 3).

Research Question 2: Linear Mixed Modeling and Fixed Effects

We explored three fixed effects—sort type, course sequence, and course topic—using three metrics: gains in edit distance to a deep sort, gains in deep pairings, and gains in deep triplets, respectively. Two fixed effects—sort type and course

TABLE 3. Average deep-problem groupings, pre and post, under both framed and unframed sort types with SEMs in parentheses

		Pairs			Triplets		
Sort type	*N*	Pre	Post	Δ	Pre	Post	Δ
Framed	412	12.2 (0.3)	13.8 (0.3)	1.6 (0.4)	6.1 (0.2)	7.4 (0.2)	1.3 (0.3)
Unframed	339	7.9 (0.3)	9.5 (0.3)	1.6 (0.4)	3.1 (0.2)	4.0 (0.2)	0.9 (0.3)

TABLE 4. Significance testing and effect size analysis of three fixed effects used in linear mixed modeling supports the framework and dosage effect propositions (column headers), but not the course topic proposition[a]

	Fixed effects							
	Sort type			Course sequence			Course topic	
	Test statistic	*p* Value	Effect size	Test statistic	*p* Value	Effect size	Test statistic	*p* Value
Change in edit distance to deep sort	$F_{1,13} = 8.27$	0.01*	$\eta^2 = 0.32$	$F_{1,13} = 3.93$	0.07[m]	$\eta^2 = 0.15$	$F_{1,13} = 0.06$	0.81
Gains in deep pairs	$F_{1,13} = 3.82$	0.07[m]	$\eta^2 = 0.16$	$F_{1,13} = 5.19$	0.04*	$\eta^2 = 0.21$	$F_{1,13} = 0.01$	0.92
Gains in deep triplets	$F_{1,13} = 2.25$	0.16	n/a	$F_{1,13} = 1.041$	0.33	n/a	$F_{1,13} = 0.004$	0.95

[a]Sort type was a significant and large predictor of change in edit distance to deep sort, but only a marginal and moderate predictor of gains in deep pairs. Likewise, course sequence was a significant and large predictor of gains in deep pairs, but a marginal and moderate predictor of changes in edit distance to deep sort. Course topic did not predict changes in any metric. Effect sizes are shown for all significant (*) and marginally significant ([m]) results; n/a = not applicable.

sequence—had significant or marginally significant effects on two metrics—edit distance to deep sort (ED_{deep}), and gains in deep pairings (Table 4). The course topic fixed effect did not explain gains in any metric. Overall, no fixed effect explained gains in deep triplets, but there were effects at finer grains of comparison. There were no discernible pairwise interactions among our grouping variables.

Gains in Edit Distance to Deep Sort

Linear mixed modeling using change in edit distance to deep sort (ΔED_{deep}) as the dependent variable showed that providing students with an expert framework (the framed sort condition; framework proposition) resulted in students making significantly greater relative gains toward the deep-feature sort (Figure 4 and Table 4). These relative gains were marginally larger in the second than the first course in the introductory sequence (dosage effect), and were unaffected by course topic (topic proposition; Table 4). There were no interactions among grouping variables for ΔED_{deep}. Course section data are given in Supplemental Table S2.

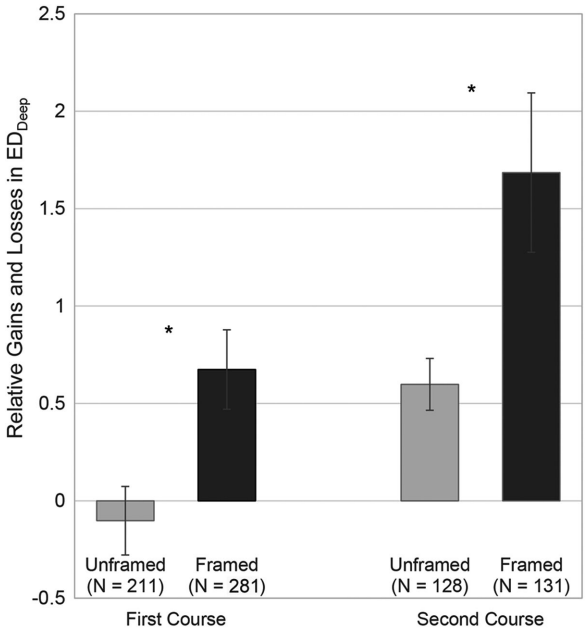

FIGURE 4. Changes in edit distance to a deep sort were significantly greater under the framed (black bars) than the unframed (gray bars) sort types ($p = 0.01$; Table 4) and marginally greater after a second course (right) than a first course (left) ($p = 0.07$; Table 4). Asterisks indicate statistically significant results.

Gains in Deep Pairs and Triplets

Results of linear mixed modeling on single problems under the framed sort condition showed that neither course sequence nor course topic predicted gains toward deep-feature sorting of the 16 single problems, either alone or in interaction with other grouping variables (all p values > 0.05).

In contrast, results of linear mixed modeling on deep-feature pairs showed that course sequence (dosage effect) significantly predicted gains toward expertise, with a marginal effect of sort type (framework proposition; Table 4) on gains. Students in their second biology course made significantly greater normalized gains on 13 of 24 deep pairs (Figure 5) than students in their first course, and the effects were large. Course topic (topic proposition) was not a significant predictor of changes in deep-feature pairings. There were no interaction effects between any pairs of grouping variables.

Results of linear mixed modeling on deep-feature triplets showed a more complex pattern. While no fixed effect explained gains in aggregated deep triplets (Table 4), two fixed effects each explained gains within two different deep-concept areas. Course sequence (dosage effect) was the only predictor of gains in deep-feature association for three of four S&F deep-feature triplets, and the effects were significant and large (all $p \leq 0.05$; all Cohen's $d \geq 0.9$). On the other hand, sort type (framework proposition) was the significant predictor of gains in deep-feature association for three of four Evol deep triplets; the effects were significant and large (all $p \leq 0.05$; all Cohen's $d \geq 1.1$). Course topic (topic proposition) was not a significant predictor of normalized gains or losses in any deep-feature triplets. There were no pairwise interactions among any of the grouping variables.

DISCUSSION

The aim of the present study was to begin to characterize expertise development among large, heterogeneous populations of introductory biology students, and our results build on previous research in two specific ways. First, we demonstrate that the BCST discerns changes in expertise over a relatively short time span among individuals in populations best characterized as novice. Second, we demonstrate that the BCST is capable of illuminating changes in expertise based upon the availability of an external deep-feature framework (captured by the sort type) and a dosage effect (captured by whether the course was first or second in the sequence).

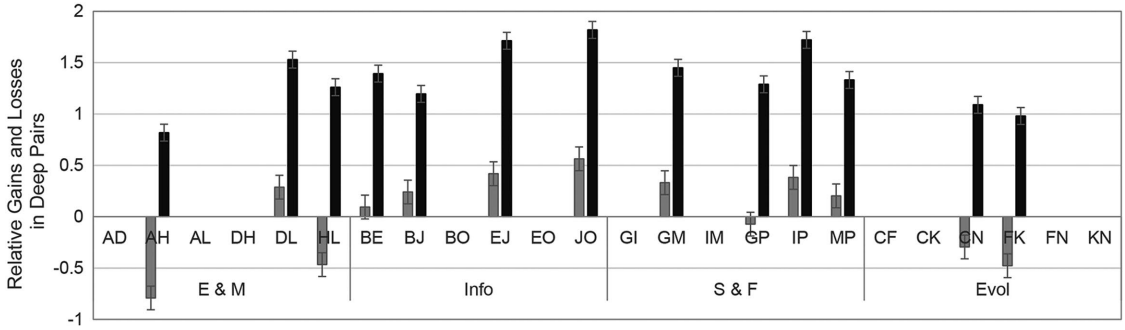

FIGURE 5. Thirteen of 24 deep-feature pairs showed significant and large gains (all $p \leq 0.05$, Cohen's $d \geq 0.8$) in deep-feature sorting by students in their second semester of introductory biology (black bars) compared with first-semester students (gray bars). For readability, only data for the 13 deep pairs showing significant differences between first and second courses are shown. Error bars are SEM.

Research Question 1. Can the BCST Detect Changes over a Single Semester?

The BCST detected population-level changes in expertise among students in their first or second semester of introductory biology. Decreases in superficial pairings, increases in deep pairings, and a decrease in unexpected problem pairings under the framed sort (Figure 3) support this claim. These data are consistent with results from other card-sorting research results (Chi *et al.*, 1981; Sanders *et al.*, 2005; Smith *et al.*, 2013), but are the first to show changes in biological expertise over a single semester of introductory biology and within a population described as predominantly novice. Moreover, cluster analyses (Supplemental Table S1) support a superficial framework based on organisms and a deep conceptual framework in biology based on core concepts from *Vision and Change* (AAAS, 2011). In other words, students were not merely slotting the biological ideas and facts they encountered into an arbitrary, externally imposed framework. The deep conceptual framework that experts build closely resembles one based upon four core concepts in *Vision and Change*.

Our results also show that novices are significantly more likely to link problems based on deep-concept principles when they have access to the deep conceptual framework and under the framed sorting condition, in which superficial and unexpected pairings both decreased, while deep pairings increased. This pattern holds for both pairs and triplets (Figure 4 and Table 3). The implication for introductory biology instruction is that students should benefit from early and frequent access to the deep conceptual framework, along with opportunities to explore, test, and explain links among facts and ideas. Interestingly, when students in our study were provided the deep-feature categories under the framed sorting condition, a small but persistent fraction of the problem pairings (11–13%, Figure 3A) were still based on superficial problem features, even after one or two semesters of introductory college biology. Thus, providing a deep conceptual framework is a necessary but insufficient precondition for expertise development. We also characterized changes in fractions of problems grouped as deep triplets. Although the overall frequencies of deep-concept triplets were low, the frequencies of deep-concept triplets increased significantly, pre to post, under both sorting tasks (Table 3). Because they represent a three-way linkage among ideas, deep-concept triplets are unlikely to occur by chance alone. Thus, deep concept may be a promising way to identify individuals or subpopulations of people who are farther along in their advancement to expertise (e.g., initiates or apprentices; Hoffman, 1996).

Our data provide strong evidence that the BCST is a robust tool for detecting changes in expertise over a single semester of introductory college biology.

Research Question 2. Which (If Any) Factors Help Explain How Relative Novices Develop Expertise over a Single Semester of Introductory Biology?

We explored the utility of three propositions for explaining the development of expertise in introductory biology students: provision of an expert framework (on the framed vs. unframed sort), a dosage effect (students in second vs. first biology course), and course topic (cellular–molecular topics vs. organismal–population topics). We found that the first two propositions helped to explain changes in expertise, but course topic did not. Specifically, we found support for the framework proposition from changes in students' edit distance to a deep sort; the results were both significant and large (Table 4). Gains in deep pairs marginally supported the framework proposition, with a moderate effect size. We found support for a dosage effect from gains in deep pairings, with a result that was significant and moderate (Table 4). Change in edit distance to deep sort marginally supported the dosage effect, also with a moderate effect size. We describe our results addressing these three propositions, implications for research, and implications for curricular design and pedagogy, in greater detail below.

Framework Proposition

Students made significant advances toward deep-feature sorting under the framed sorting condition. Supporting this claim is the change in edit distance to deep sort, which was significantly larger for the framed sort type than the unframed sort type (Table 4 and Figure 4). Course sequence (whether first or second) only marginally explained change in edit distance to deep sort. However, students' advancement toward biological expertise occurs alongside other phenomena. For example, unexpected pairings accounted for a remarkably robust 30–40% of all problem pairings (Figure 3), consistent with data reported by Smith and colleagues (2013). Also, edit distance to a

deep-feature sort changed independently of edit distance to a superficial-feature sort (unpublished data).

Students made gains in deep-feature sorting under both sort types (Figures 3 and 4). When provided the deep conceptual framework on the framed sort, students were able to sort deep-feature problem pairs marginally significantly more often than on the unframed task (Table 4), in which they had to generate the categories from their existing frameworks themselves.

Taken together, these lines of evidence support a claim that it is important for novices to have access to a deep conceptual framework before they are able to generate it themselves. Throughout their experiences in our courses, it is likely that our novice population is testing, forming, and breaking links among ideas as they build and improve their nascent conceptual frameworks. Our results suggest that consistent access to the deep conceptual framework, along with instructor- and student-generated explanations of these linkages, might be critically important to novice students early in their journeys toward expertise. One question prompted by these results is how concepts and ideas interact with one another in the novice's evolving conceptual framework. There is empirical support for both the superficial- and deep-feature organization frameworks of novices and experts, respectively, but the dynamics and mechanisms that help explain how a superficial framework evolves into a deep conceptual framework are unknown.

Dosage Effect

When we examined course sequence as a fixed effect, we found that it had a significant effect on gains in deep pairs, and a marginally significant effect on ED_{deep}. Students in their second biology course had greater normalized gains in 13 of 24 deep-concept pairs (Table 4 and Figure 5): 3/6 deep pairs in E&M, 4/6 deep pairs in both Info and S&F, and 2/6 deep pairs for Evol. Although course sequence had a marginal effect on changes in edit distance to deep sort, the effect was significant on the subset of students completing the framed sort (Figure 4), highlighting the critical role of access to an expert framework as a pre condition for linking facts and ideas. These results may be explained by how developing experts use their experiences with facts and ideas to filter and conditionalize what they pay attention to in the future (Gobet and Simon, 1996). As students progress through the curriculum, the time they spend working with facts and ideas should increase. Because student experiences shape their abilities to filter and conditionalize their knowledge, it may be critically important to their development of expertise that such experiences are meaningfully grounded in deep concepts and scientific practices.

Topic Proposition

There was no difference in the effect of course topics on student changes in expertise among any of the three metrics we used, alone or in conjunction with the other two fixed effects. This set of evidence, at least, runs counter to beliefs of some instructors. However, we recognize that course topic is a very coarse grain size to apply to courses that both include a mixture of concrete and abstract, or descriptive, hypothetical, and theoretical topics, as in Lawson and colleagues' (2000) framework. At this university, instructors are also free to structure their individual courses as they believe will help students' learning and to sequence a broad set of topics as they choose. It is possible that particular topics within courses themselves may influence whether students advance in expertise, and by how much. Because we were unable to test this hypothesis at a finer grain, it deserves further scrutiny.

Caveats and Synthesis

The present study did not include a control group of students who were not enrolled in any biology course but who completed the same problem-sorting task at the beginning and end of an academic semester. Hence, one question that remains is whether people can develop expertise through observational or vicarious learning: reading about biological topics for pleasure, watching engaging videos, or visiting museums, for example, instead of being enrolled in an academic course focused on biology. We would predict that students not exposed to biological concepts or the deep conceptual framework will not gain in biological expertise, but this is an open question for investigation. We also assumed in the present study that a dosage effect corresponded to exposure time rather than completeness of the dose of biology. Hence, another remaining question is whether and how instructional alignment with deep concepts in biology affects students' expertise development. Finally, a plausible alternative explanation of the significance of course sequence on students' developing expertise is that student populations in second-semester biology courses may be more self-selective in their enrollment, explaining advances in expertise through greater motivation, attention, aptitude, or some combination of those factors that this investigation did not measure. With these caveats in mind, we offer the following synthesis.

Interestingly, while our results supported two of our three propositions, different metrics supported each to a greater or lesser extent (Table 4), most likely due to the heterogeneity of our student populations (Tables 1 and 2). Sort type—the availability of an expert conceptual framework—was a significant predictor of change in edit distance, a variable that describes a student's ability to match a problem with its deep concept in a provided deep conceptual framework. On the other hand, course sequence—in our study, a dosage effect—predicted gains in deep pairs, a variable that measures students' abilities to actually connect and situate two or more ideas within their deep concept. These data support both the framework proposition and the dosage effect as potential explanations for advancement toward expertise in populations of putative novices. Seeing how experts organize concepts and ideas is important, but it doesn't replace the experiences of novices forming and testing links among facts and ideas themselves—an effect that accumulates over time.

Data from deep-concept triplets were equivocal. Although aggregated triplet gains support none of our three hypotheses (Table 4), at a finer grain, conceptual triplet groups did show some significant trends. Three of four S&F triplets differed significantly between students in their first or second courses (e.g., course sequence), whereas three of four Evol triplets differed significantly between students in the framed or the unframed sort (e.g., the framework proposition). If our propositions are supported, then perhaps it is more important for students to be exposed over time to some biological concepts, such as for S&F, whereas, for other concepts, such as Evol, it may be more

important for students to have access to the expert framework as they build and adapt their own templates. In biology, systems show emergent properties based on multiple links among system components that can wax and wane over time. Expertise has properties of an emergent system as well, with multiple links among ideas and concepts that are conditionalized to a particular context. In our data, none of the individual BCST individual problems differed between populations delimited by any of our grouping variables. In contrast, 13 of 24 deep pairs (Figure 5) and six of 16 deep-concept triplets (unpublished data) differed significantly between populations. Together, these are some intriguing glimpses into the emergent organization of developing expertise. Curricula and pedagogies that allow students to grapple with interactions among multiple ideas could introduce them to this important feature of biological systems at the same time that the process is helping students to build their expertise.

We believe that our results lend further support to the importance of planning and organizing curricula and course activities around the core concepts of our discipline. Conceptual organization may be especially important with populations of novices in introductory courses, and particularly the first course in a sequence. Instructors sometimes bemoan students' inabilities to make connections among what they have learned or to extend their thinking. Our results suggest that one way to meet aspirations of "undergraduates…integrat[ing] concepts across levels of organization and complexity and…synthesiz[ing] and analyz[ing] information" (AAAS, 2011, p. ix) is by providing and explaining the deep conceptual framework, by structuring our course curricula around a the deep concepts of our discipline, and by allowing students to grapple with constructing their own explanations of and links between facts and ideas. This approach contrasts that of many biology texts (Koulaidis and Tsatsaroni, 1996), curricula and pedagogical practices (Goffe and Kauper, 2014), and assessments (Momsen *et al.*, 2010). Feldon (2006) suggested a curricular approach based on cognitive task analysis—that is, systematically breaking down the steps in a process of understanding the kinds of complex cognitive tasks we ask our students to complete. The provision of an expert framework and a dosage effect together may help explain advances in expertise, but their effects may vary based on where a student is on the novice–expert continuum. If deep-concept triplets are indeed a reliable diagnostic of more-developed novices, then these results point to another question worth investigating: whether and how different factors are more or less important for expertise development for novices at different stages of their development.

FUTURE RESEARCH

This study is an example of how we tested theories from cognitive science "in the wild." We suggest three areas of investigation for the immediate future: exploring the effects of using different problems in the tool itself, expanding to different student populations, and accounting for different learning environments, curricular arrangements, and student motivation. First, the BCST's current design consists of 16 problems. One opportunity lies in investigating whether unknown aspects of the problems themselves or the superficial- and deep-concept combination represented in each problem influences BCST results, perhaps by choosing and vetting different problems to represent the same superficial–deep combinations. Next, for the purposes of the present study, we have considered students as a single, heterogeneous population. Hence, another set of questions centers on investigating whether different demographic subgroups develop expertise differently. One testable proposition arising from our results is that courses focusing on a deep conceptual framework and upon students forming and testing links among facts and ideas should lead to greater gains in expertise. Finally, the higher-educational settings of our laboratories and classrooms are themselves heterogeneous: many include incidental or deliberate social interactions among students and instructors, all of whose motivation to master biological and social issues may vary. We also assumed, for this investigation, that all biology courses were equivalent in their effect on students enrolled in the course. Another potentially productive area of research, then, is the influence of environmental factors, including social interactions among students and between students and instructors; structural factors, such as the degree to which courses are conceptually driven; and the role of motivation on the development of expertise within individuals and among groups. Together, these theory-driven investigations have the potential to influence how we help novices in their journeys toward expertise in our biology classrooms and beyond.

ACKNOWLEDGMENTS

We thank K. D. Tanner and E. D. Combs for making the BCST and analysis tools available to us and for productive conversations about their development and vetting of the BCST. E. D. Combs provided technical support for applying the Python scripts he developed. W. Ma of MSU's Center for Statistical Training and Consulting provided statistical consultation. K. D. Tanner, C. M. Trujillo, and S. Wolf provided many helpful comments on an early draft of the paper. J. Abatie, K. Gordon, and T. Orlando helped with data entry. Finally, we are grateful to the introductory biology instructors and students who participated in this study and made this research possible. This work was supported under National Science Foundation DUE 08172224, Faculty Institutes for Reforming Science Teaching IV, D.E.-M. (principal investigator).

REFERENCES

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

Campbell M, Heyer LJ, Paradise CJ (2014). Integrating Concepts in Biology, 1st ed. www.trunity.com/trubook-integrating-concepts-in-biology-by-campbell-heyer-paradise.html (accessed 6 December 2016).

Chase WG, Simon HA (1973). Perception in chess. Cogn Psychol 4, 55–81.

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

de Bruin ABH, Kok EM, Leppink J, Camp G (2014). It might happen in the very beginning: reply to Ericsson. Intelligence 45, 107–108.

deGroot AD (1965). Thought and Choice in Chess, The Hague, Netherlands: Mouton.

Deibel K, Anderson R, Anderson R (2005). Using edit distance to analyze card sorts. Expert Syst 22, 129–138.

Ericsson KA, Krampe RT, Tesch-Römer C (1993). The role of deliberate practice in the acquisition of expert performance. Psychol Rev 100, 363–406.

Feldon DF (2006). The implications of research on expertise for curriculum and pedagogy. Educ Psychol Rev 19, 91–110.

Fincher S, Tenenberg J (2005). Making sense of card sorting data. Expert Syst 22, 89–93.

Forster KI, Davis C (1984). Repetition priming and frequency attenuation in lexical access. J Exp Psychol Learn Mem Cogn 10, 680–698.

Glaser R, Chi MTH (1988). Overview. In: The Nature of Expertise, ed. MTH Chi, R Glaser, and M Farr, Hillsdale, NJ: Erlbaum, xv–xxviii.

Gobet F (1998). Expert memory: a comparison of four theories. Cognition 66, 115–152.

Gobet F (2005). Chunking models of expertise: implications for education. Appl Cogn Psychol 19, 183–204.

Gobet F, Simon HA (1996). Recall of random and distorted chess positions: implications for the theory of expertise. Mem Cognit 24, 493–503.

Goffe WL, Kauper D (2014). A survey of principles instructors: why lecture prevails. J Econ Educ 45, 360–375.

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.

Hambrick DZ, Oswald FL, Altmann EM, Meinz EJ, Gobet F, Campitelli G (2014). Accounting for expert performance: the devil is in the details. Intelligence 45, 34–45.

Hoffman RR (1996). How can expertise be defined? Implications of research from cognitive psychology. In: Exploring Expertise, ed. R Williams, W Faulkner, and J Fleck, Edinburgh, UK: University of Edinburgh Press, 81–100.

Koulaidis V, Tsatsaroni A (1996). A pedagogical analysis of science textbooks: how can we proceed? Res Sci Educ 26, 55–71.

Lawson AE, Alkhoury S, Benford R, Clark BR, Falconer KA (2000). What kinds of scientific concepts exist? Concept construction and intellectual development in college biology. J Res Sci Teach 37, 996–1018.

Maher JM, Markey JC, Ebert-May D (2013). The other half of the story: effect size analysis in quantitative research. CBE Life Sci Educ 12, 345–351.

McCauley R, Murphy L, Westbrook S, Haller S, Zander C, Fossum T, Sanders K, Morrison B, Richards B, Anderson R (2005). What do successful computer science students know? An integrative analysis using card sort measures and content analysis to evaluate graduating students' knowledge of programming concepts. Expert Syst 22, 147–159.

Medin DL, Lynch EB, Solomon KO (2000). Are there kinds of concepts? Annu Rev Psychol 51, 121–147.

Michael J (2007). What makes physiology hard for students to learn? Results of a faculty survey. Adv Physiol Educ 31, 34–40.

Momsen JL, Long TM, Wyse SA, Ebert-May D (2010). Just the facts? Introductory undergraduate biology courses focus on low-level cognitive skills. CBE Life Sci Educ 9, 435–440.

National Research Council (NRC) (1999). Transforming Undergraduate Education in Science, Mathematics, Engineering, and Technology, Washington DC: National Academies Press.

NRC (2000). How People Learn: Brain, Mind, Experience, and School, Washington DC: National Academies Press.

Rutherford FJ, Ahlgren A (1991). Science for All Americans, Oxford University Press.

Sanders K, Fincher S, Bouvier D, Lewandowski G, Morrison B, Murphy L, Petre M, Richards B, Tenenberg J, Thomas L, *et al.* (2005). A multi-institutional, multinational study of programming concepts using card sort data. Expert Syst 22, 121–128.

Schoenfeld AH, Herrmann DJ (1982). Problem perception and knowledge structure in expert and novice mathematical problem solvers. J Exp Psychol Learn Mem Cogn 8, 484–494.

Shanteau J (1992). The psychology of experts: an alternative view. In: Expertise and Decision Support, ed. G Wright and F. Bolger, New York: Plenum, 11–23.

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

Solomon KO, Medin DL, Lynch E (1999). Concepts do more than categorize. Trends Cogn Sci 3, 99–105.

Wolf SF, Dougherty DP, Kortemeyer G (2012). Rigging the deck: selecting good problems for expert-novice card-sorting experiments. Phys Rev Spec Topics Phys Educ Res 8, 020116.