# Using the Biology Card Sorting Task to Measure Changes in Conceptual Expertise during Postsecondary Biology Education

**Sarah A. Bissonnette,[†] Elijah D. Combs,[†] Paul H. Nagami,[‡§] Victor Byers,[§] Juliana Fernandez,[§] Dinh Le,[§] Jared Realin,[§] Selina Woodham,[§] Julia I. Smith,[§] and Kimberly D. Tanner[†]***

[†]Department of Biology, San Francisco State University, San Francisco, CA 94132; [‡]Department of Biology, Laney College, Oakland, CA 94607; [§]Division of Math and Science, Holy Names University, Oakland, CA 94619

## ABSTRACT

**While there have been concerted efforts to reform undergraduate biology toward teaching students to organize their conceptual knowledge like experts, there are few tools that attempt to measure this. We previously developed the Biology Card Sorting Task (BCST), designed to probe how individuals organize their conceptual biological knowledge. Previous results showed the BCST could differentiate between different populations, namely non–biology majors (NBM) and biology faculty (BF). In this study, we administered the BCST to three additional populations, using a cross-sectional design: entering biology majors (EBM), advanced biology majors (ABM), and biology graduate students (BGS). Intriguingly, ABM did not initially sort like experts any more frequently than EBM. However, once the deep-feature framework was revealed, ABM were able to sort like experts more readily than did EBM. These results are consistent with the conclusion that biology education enables advanced biology students to use an expert-like conceptual framework. However, these results are also consistent with a process of "selection," wherein students who persist in the major may have already had an expert-like conceptual framework to begin with. These results demonstrate the utility of the BCST in measuring differences between groups of students over the course of their undergraduate education.**

## INTRODUCTION

Reform efforts currently underway in undergraduate biology education emphasize teaching students to think like biologists and discourage the traditional single-minded focus on content coverage. The *Vision and Change in Undergraduate Biology Education* report outlined a number of suggestions for reforming undergraduate biology education so that, as a result of their education, students with a biology degree would "become adept at making connections among seemingly disparate pieces of information, concepts, and questions" and to "ensure that the biology we teach reflects the biology we practice" (American Association for the Advancement of Science [AAAS], 2011, pp. 3 and viii). As a result of their biology education, then, biology students should have greater conceptual expertise in biology. Here, we define conceptual expertise as a more expert-like organization of conceptual biology knowledge. To assess their efforts to reform undergraduate biology education, departments, programs, and instructors need a way to measure changes in conceptual expertise. How could such conceptual expertise be measured?

Early attempts to investigate differences in disciplinary conceptual expertise between novices and experts came from physics education research. In a key paper, Chi and colleagues tasked physics graduate students—putative experts—and physics undergraduate students—putative novices—to sort 24 introductory physics problems based on "similarities of solution" (Chi *et al.*, 1981). They found that their putative

**A. Malcolm Campbell,** *Monitoring Editor*

Submitted September 8, 2016; Revised December 13, 2016; Accepted December 18, 2016

CBE Life Sci Educ March 1, 2017 16:ar14

DOI:10.1187/cbe.16-09-0273

*Address correspondence to: Kimberly D. Tanner (kdtanner@sfsu.edu).

experts tended to sort the problems based on underlying physics principles, or so-called "deep structures" (e.g., Newton's second law), while putative novices tended to sort the problems based on so-called "surface features" of the question, such as whether inclined planes were used in the problem. The Chi *et al.* study, as well as many subsequent studies, suggested that novices and experts could be differentiated by how they performed on a problem-sorting task (Weiser and Shertz, 1983; Smith, 1990; Mason and Singh, 2011; Krieter *et al.*, 2016).

Given the dearth of tools to specifically measure the organization of disciplinary conceptual knowledge in biology, we previously developed the Biology Card Sorting Task (BCST; Smith *et al.*, 2013). We modeled the BCST on the task in the Chi *et al.* (1981) study, modifying the task in several key ways to enable quantitative measurements between different and large populations. First, unlike previous studies, the BCST was based on a hypothesis-driven card stimulus set. Sixteen biology problems were chosen for the BCST, and each card represented one of four hypothesized deep features—core biology concepts— and one of four hypothesized surface features—organism types. The four hypothesized deep features were four of the core concepts for biological literacy as proposed in *Vision and Change* (AAAS, 2011): evolution; structure and function; information flow, exchange, and storage; and pathways and transformations of energy and matter. The four hypothesized surface features were the organisms mentioned in the problems: humans, insects, plants, and microbes. Second, the use of a hypothesis-driven card sort allowed for the calculation of several quantitative metrics, including the percentage of deep, surface, or unexpected card pairs present in each individual's sort, as well as edit distances, which are the number of cards that would have to be moved to transform an individual's sort into either an exact hypothesized deep-feature or an exact hypothesized surface-feature sort. This improved approach to card sorting has already been applied in chemistry to investigate the development of domain-specific conceptual expertise in undergraduate chemistry majors (Krieter *et al.*, 2016).

Using the novel BCST, we previously demonstrated that it could be used to differentiate between putative biology novices—non–biology majors (NBM)—and putative biology experts—biology faculty (BF). For example, NBM generated fewer deep-feature card pairs and had card sorts that were much closer to the hypothesized surface-feature card sort compared with BF (Smith *et al.*, 2013). Interestingly, 100% of BF reported using deep-feature rationales to organize their cards, while NBM reported using a variety of different rationales to sort their cards, including but not limited to, surface-feature rationales. Intriguingly, similar findings—that while experts tend to converge on one of a small number of methods for organizing their knowledge, novices tend to be much more varied in how they organize their knowledge—have been reported for a genetics card-sorting task (Smith, 1990), recently developed chemistry card-sorting tasks (Irby *et al.*, 2016; Krieter *et al.*, 2016), and a physics card-sorting task subsequent to the work of Chi *et al.* (1981; Mason and Singh, 2011). This ability to reveal the frameworks individuals use to organize their disciplinary knowledge is a significant advantage of card-sorting tasks compared with other assessment approaches, because the organization of one's knowledge, rather than the presence, absence, or accuracy of the content knowledge itself, is thought to be key to developing expertise (Ambrose *et al.*, 2010). Furthermore, having a more expert-like organization of conceptual knowledge has been shown to be associated with increased problem-solving abilities (Eylon and Reif, 1984; Hardiman *et al.*, 1989).

The BCST was established as a tool that could address a variety of important questions. For example, to what extent are students who are graduating from biology departments organizing their knowledge any differently from entering students? And among our entering students, are there any detectable differences in how two populations—students who are strongly expressing an interest in biology and students who are nonmajors in a biology general education course—organize their conceptual biology knowledge? Having demonstrated the utility of the BCST in differentiating between biology novices and biology experts, we investigate these questions here by administering the BCST to student populations with various levels of formal biology education experience: entering biology majors (EBM), advanced biology majors (ABM), and biology graduate students (BGS). The following two overarching research questions drove the research design and data analyses. First, to what extent do student populations at different levels of formal biology education organize their biology knowledge differently? Second, to what extent can the BCST be used to uncover unhypothesized frameworks used by biology students to organize their biology knowledge?

## METHODS

The BCST is an assessment tool designed to measure conceptual expertise in biology (Smith *et al.*, 2013). The BCST asks subjects to explore associations among 16 biology problems (presented on cards) and by doing so reveal how they connect or do not connect core biological concepts (Smith *et al.*, 2013). Previous work has shown consistent and significant differences in how putative experts (BF) and putative novices (NBM) sort the biology problems in the BCST (Smith *et al.*, 2013). In this study, we expanded the scope of participant populations to include putative intermediate levels of biological expertise. Below we describe the populations sampled, the implementation of the task, and the analytic approaches used to quantify sorting differences within and across populations with regard to: constructed card groupings, constructed card-group names, and responses to reflective prompts.

### Recruitment and Participant Populations

Participants were recruited from the students and faculty of a large, urban university with more than 25,000 undergraduates (1800 biology majors and 5000 students enrolled in biology courses per term) and ~40 faculty in biology who are active in research, as well as teaching, and represent a wide breadth of subdisciplines from the molecular to the ecological scale (Table 1).

*Non–biology majors* (NBM) were recruited on the first day of their laboratory section for a general education course in biology. We hypothesized that these nonmajors would have the greatest interest in and understanding of biology among the wider population of NBM on campus, and thus we thought that querying this population would lessen the chance of artificially inflating the differences between NBM and other populations. Each student in the course completed the tasks associated with this study as part of their course curriculum, but only those

TABLE 1. Participant population

| Participant type | Participation rate | Sample size | Female participants | Participants of color |
|---|---|---|---|---|
| Non–biology majors | 89% | 101 | 55% | 56% |
| Entering biology majors | 90% | 185 | 70% | 70% |
| Advanced biology majors | 97% | 109 | 59% | 74% |
| Biology graduate students | 18% | 29 | 59% | 32% |
| Biology faculty | 69% | 23 | 26% | 48% |

identified as NBM were included in the study ($n = 101$). These data were included in a previous publication (Smith *et al.*, 2013) and are reproduced here for ease of comparison.

*Entering biology majors* (EBM) were recruited on the first day of their laboratory section for an introductory biology course. This course is a one-semester component of a yearlong introductory biology series. Only subjects from this course who were declared biology majors were included in the analyses ($n = 185$). Seventy-five percent of this class was made up of students who were more than five semesters away from graduation (i.e., students who would traditionally be considered freshmen and sophomores) as determined by the students' self-reported anticipated year of graduation.

*Advanced biology majors* (ABM) were recruited on the last day of their genetics class; this is the most advanced biology class required of all undergraduate concentrations in biology. Only subjects from the genetics course who were declared biology majors were included in the analyses ($n = 109$). Ninety-five percent of this class was made up students who were within four semesters of graduation (i.e., students who would traditionally be considered juniors and seniors) as determined by the students' self-reported anticipated year of graduation.

*Biology graduate students* (BGS), who were graduate students pursuing a master's degree in biology with concentrations in cell/molecular biology, ecology/systematic biology, conservation biology, physiological/behavioral biology, and microbiology, were recruited through email. The tasks associated with the study were administered in a small-group setting during a single semester ($n = 29$).

*Biology faculty* (BF), who were tenured and tenure-track members of the biology faculty were recruited by email. Tasks associated with the study were administered individually during a single semester by a member (J.I.S.) of the research team who was not known to them ($n = 23$). These data were included in a previous publication (Smith *et al.*, 2013) and are reproduced here for ease of comparison.

| | Hypothesized Deep Features | | | |
|---|---|---|---|---|
| Hypothesized Surface Features | Evolution and 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 |
| Plant | K | D | J | I |
| Insect | H | F | B | M |
| Human | N | L | O | G |
| Microorganism | C | A | E | P |

FIGURE 1. Biology Card Sort Task (BCST) design. Columns represent the four different hypothesized deep features of biology, the rows represent the four hypothesized surface features of biology. Each letter represents one of the 16 biology problems included in the BCST. Each problem was chosen to exemplify one hypothesized deep feature and one hypothesized surface feature.

## Task Conditions

We used the previously described hypothesis-driven card set (Figure 1) to probe how populations organized biological concepts. Each individual biology problem was written on a separate card. Participants were asked to read the questions but not to attempt to solve the problems. Participants were told that the task was not intended as a test and that there were no right or wrong ways to sort the cards. The card-sorting task was performed under two conditions (Figure 2): first the unframed task condition, and then the framed task condition (described below).

*Unframed Sorting-Task Condition.* Under the unframed condition (Figure 2, left-hand column), participants were asked to consider what they knew about biology and to sort the problems into groups representing common *underlying biological principles*. No further instructions were given as to what was meant by the phrase "underlying biological principles." Subjects were told that they must generate more than one group and fewer than 16 groups of cards and that any problem could be a member of only one group. Each group constructed was to be given a name that described the commonality that caused individuals to group those problems together. Subjects recorded their constructed groups, the names of those groups, their start and stop times, and the number of groups they created on a form provided by the researchers. When subjects were finished sorting in the unframed condition, they were asked to respond to two reflective prompts that probed the reasoning behind their card groupings and group names: 1) *Describe why you grouped certain problems together and give an example of your reasoning*, and 2) *How did you decide on the names of your groups?*

*Framed Sorting-Task Condition.* Once subjects had completed the unframed task, they were asked to sort the problems a second time (Figure 2, right-hand column). The framed condition was employed to determine the extent to which subjects could use the hypothesized deep-feature framework when specifically prompted

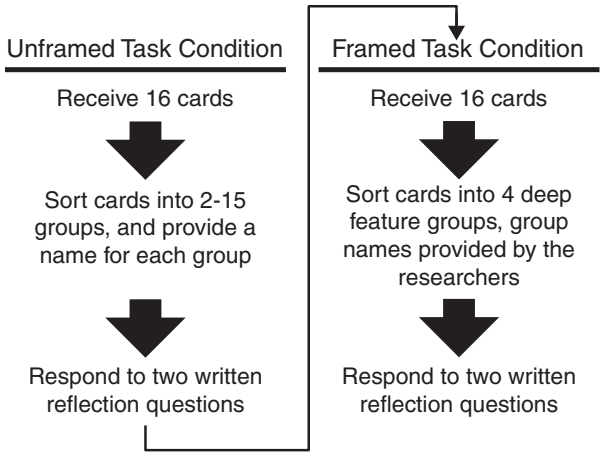


**FIGURE 2. Protocol for BCST administration. The BCST is conducted in two phases: the unframed phase always precedes the framed phase.**

with those concepts. In the framed sorting-task condition, participants were asked to sort the same 16 problems again, but this time into four groups that had been preassigned the following names by the researchers: 1) Evolution by Natural Selection in Living Systems, 2) Pathways and Transformations of Energy and Matter in Living Systems, 3) Storage and Passage of Information about How to Build Living Systems, and 4) Relationships between Structure and Function in Living Systems. Subjects recorded their constructed groups and their start and stop times on a form provided by the researchers. When subjects were finished sorting in the framed condition, they were asked to respond to two reflective prompts: 1) *Which if any of the problems was difficult to assign to one of the 4 categories and why? Please list all that apply*; and 2) *Now that you have completed 2 card-sorting activities, which group names do you prefer: the group names that you created or the group names given to you by the researchers or neither? Please explain your answer*. Analyses of the answers to these two prompts goes beyond the scope of the current research report and were not systematically analyzed.

After completing both of the task conditions, subjects were asked to respond to a variety of demographic questions regarding themselves and their educational backgrounds. Only those participants who completed all tasks as directed by the researchers were included in the study. This research was approved both by the committee for the protection of human subjects at San Francisco State University (Protocol #X10-036) and the Institutional Review Board for the Protection of Human Subjects at Holy Names University (Protocol Title: "Investigation of How Novices and Experts Structure Knowledge of Fundamental Biological Principles").

### Analyses and Comparison of Constructed Card Groupings

Subjects may have organized their cards based upon hypothesized surface features (organism type), hypothesized deep features (core biological concepts), or some other unexpected organization. To quantify how similar the card groupings generated by participants were to our hypothesized groupings (Figure 1), we used two quantitative metrics to describe each individual's sort: percent card pairings and edit distance.

### Percent Card Pairings

Percent card pairings measured the degree to which the cards grouped by a participant generated pairings predicted as surface-feature pairings, deep-feature pairings, or unexpected pairings. For example, in the card group {CDK}, one card pair—CK—belongs to the hypothesized deep-feature group, Evolution by Natural Selection in Living Systems (Figure 1). Another pair—DK—belongs to the surface-feature group, Plant (Figure 1). The final card pair—CD—represents an unexpected pairing; it belongs to neither the hypothesized surface- nor deep-feature groupings (Figure 1). The BCST contains 136 possible card pairs, 24 of which are hypothesized deep-feature pairings, 24 of which are hypothesized surface-feature pairings, and 88 of which are unexpected pairings. If a participant generated a group with only a single card, then that individual card was treated as an unexpected pairing. Percentages of deep-feature, surface-feature, and unexpected card pairings were calculated for each participant (using a data entry Python script [Python Software Foundation, 2011] written by the researchers and freely available by contacting the corresponding author) by identifying all the card pairs within each card group for each group generated by the participant and dividing by the total number of pairs produced by that participant. Percentages of deep-feature pairings, surface-feature pairings, and unexpected pairings were averaged across individuals in each participant population for both the unframed and framed conditions and were then compared across populations.

### Edit Distance

A second analytical approach was used to quantify and compare sorting results. Edit distance (Deibel *et al.*, 2005) is defined as the minimum number of card moves needed to turn an individual's card sort into either an exact hypothesized surface-feature sort or an exact hypothesized deep-feature sort (Figure 1). Units of ED are "necessary card moves." An exact hypothesized deep-feature sort would have an ED-Deep of 0 necessary card moves and ED-Surface of 12 necessary card moves. Similarly, an exact hypothesized surface-feature sort would have an ED-Deep of 12 necessary card moves and ED-Surface of 0 necessary card moves. EDs were calculated using a Python script written by the researchers that used Clapper's *munkres* implementation of the Hungarian method (Kuhn, 2010; Clapper, 2008). Using this approach, an ED from the exact hypothesized surface-feature sort (ED-Surface) and an ED from the exact hypothesized deep-feature sort (ED-Deep) were calculated for each individual card sort and averaged across populations. These were then compared across populations for each task condition.

### Analyses and Comparison of Constructed Card-Group Names

Scoring rubrics were developed to determine the extent to which group names given by participants in the unframed condition employed language related to either the hypothesized deep features or surface features. (Examples of included and excluded terms for deep-feature group names are given in Supplemental Table S1.) The scoring rubrics were revised using blind coding of subsets of the data until at least 90% interrater reliability was achieved. Group names given by participants that did not match the hypothesized features were

not included in this analysis. The percentages of participants in each population who gave group names similar to the hypothesized features were calculated and then compared across populations.

### Analysis and Comparison of Responses to Reflective Prompts

Analysis of the responses to the posttask reflection questions administered after the unframed task condition revealed that many participants used rationales that were unrelated to the deep features or surface features of the cards. Therefore, a grounded theory approach was taken to analyze the responses to the reflective prompts. Then, a scoring rubric was developed to assess the prevalence of sorting strategies based upon the most common rationales, which included surface features, deep features, and curriculum-based rationales (Table 2). This rationale rubric was used to analyze all the responses given by participants to the reflection questions given after the unframed sort. The prevalence of several other sorting strategies was quantified (rationales mentioning using non–surface feature key words from the cards, rationales that make reference to a coherent strategy without making reference to a hypothesized strategy, and rationales that were too vague to accurately categorize); however, all these were prevalent in less than 20% of any population and are not shown here.

### Comparative Statistical Analyses

Two-tailed Student's *t* tests were used to compare the average percent card pairing and average ED measures between participant populations within a task condition. Only populations that were "adjacent" in biology experience were compared with each other, as these were the comparisons that were of the greatest interest, and to maintain statistical power. That is, NBM were compared with EBM, EBM were compared with ABM, ABM were compared with BGS, and BGS were compared with BF. In comparing the average percent card pairings, we considered the unframed and framed conditions to be separate families of tests. We performed four tests per card pairing (one between each adjacent populations as described above) for each of the three types of card pairings (deep, surface, and unexpected pairs) for a total of 12 tests in the unframed condition and 12 tests in the framed condition. For both the unframed and framed conditions, to correct for multiple comparisons, a Bonferroni-adjusted alpha level of 0.004 (0.05/12) was used. Comparisons of the results for a single participant population between the two task conditions—unframed and framed—were also analyzed. This was considered a separate family of tests with 15 tests, three (percent surface, percent deep, and percent unexpected pairs) for each of the five populations. In this case, to correct for multiple comparisons, a Bonferroni-adjusted alpha level of 0.003 (0.05/15) was used.

**TABLE 2. Rubric for and analysis of card-sorting strategy explanations[a]**

| Surface-feature rationale | | Sample quote |
|---|---|---|
| Non–biology major ($n$ = 101) | 37.6% | "I grouped them together based on what kind of organism they pertained to." |
| Entering biology major ($n$ = 185) | 33.0% | "An example of how I grouped my cards is I grouped all the cards that mentioned insects and put those together in one group." |
| Advanced biology major ($n$ = 109) | 32.1% | "I grouped the problems together based on the organism or group that the question related to." |
| Biology graduate student ($n$ = 29) | 13.8% | "Topics relating to microorganisms I grouped as microbiology" |
| Biology faculty ($n$ = 23) | 8.7% | "Others are united by kind of organism. (D,I; P,M,G)" |
| **Deep-feature rationale** | | **Sample quote** |
| Non–biology major ($n$ = 101) | 22.8% | "I grouped problems together that dealt with DNA or genetics of organisms." |
| Entering biology major ($n$ = 185) | 37.8% | "I grouped based on concept of each question. For example anything involving metabolism or use of energy I grouped as metabolism." |
| Advanced biology major ($n$ = 109) | 32.1% | "My biggest group is structure/function. All of the questions relate to how something is organized or what it is made up of and how that relates to its function." |
| Biology graduate student ($n$ = 29) | 48.3% | "I tried to figure out the underlying themes. For example, A, D, and L were different questions about cellular respiration and metabolism so I put them under the group "cellular energetics." |
| Biology faculty ($n$ = 23) | 100% | "I looked for similar characteristics among cards. An easy one were the cards selected for "basic structural characteristics of organisms" G, I, M, P- which were descriptive of structures and implicit or explicit implications of process." |
| **Curricular rationale** | | **Sample quote** |
| Non–biology major ($n$ = 101) | 6.9% | "It was pretty easy, almost like chapters of a textbook. |
| Entering biology major ($n$ = 185) | 10.3% | "I tried thinking about big topics in biology and what is usually grouped together in lectures." |
| Advanced biology major ($n$ = 109) | 28.4% | "I then looked at what each problem was asking and sorted based on what kind of bio class I could see having discussed such problem or may have such a problem on a test." |
| Biology graduate student ($n$ = 29) | 48.3% | "I primarily grouped them according to undergraduate Biology course titles" |
| Biology faculty ($n$ = 23) | 39.1% | "Maybe in what section of a book (intro) I'd find it, or what course, based on concepts I think cards reference." |

[a]NBM and BF data are reprinted from Smith *et al.* (2013). Adjacent populations with differences significant to a Bonferroni-adjusted $p < 0.0125$ (see *Methods*) are denoted with a vertical bracket and a number. The numbers correspond to the following statistical values: [1]$\chi^2 = 6.8$, $df = 1$, $p = 0.0093$; [2]$\chi^2 = 16.7$, $df = 1$, $p < 0.0001$; [3]$\chi^2 = 16.0$, $df = 1$, $p < 0.0001$.

In comparing the average ED-Surface and ED-Deep, we considered the unframed and framed conditions to be separate families of tests. We performed four tests for ED-Surface (one between each adjacent populations) and four tests for ED-Deep for a total of eight tests in the unframed condition and eight tests in the framed condition. Here, to correct for multiple comparisons, a Bonferroni-adjusted alpha level of 0.006 (0.05/8) was used. Comparisons of the results for a single participant population between the two task conditions—unframed and framed—were also analyzed. This was considered a separate family of tests with 10 tests, two for each population (the difference in ED-Surface between the framed and unframed conditions, and the difference in ED-Deep between the framed and unframed conditions). In these cases, to correct for multiple comparisons, a Bonferroni-adjusted alpha level of 0.005 (0.05/10) was used.

Pearson's chi-square tests were used to compare the prevalence of group names and specific card-sorting strategies used by different participant populations. Each group name or rationale was composed of four tests, one for each adjacent pair of populations. In these cases, to correct for multiple comparisons, a Bonferroni-adjusted alpha level of 0.0125 (0.05/4) was used.

To normalize for differences in the size of particular participant populations, we present all variances as an SEM. All statistical comparisons were generated using JMP version 12.1.0 (SAS Institute, 2015).

## RESULTS

In this study, we employed the BCST to investigate how organization of conceptual knowledge in biology may or may not differ among students with various amounts of formal biology education. As described previously (Smith *et al.*, 2013), the BCST yields multiple sources of data for analysis. Below, we describe the three new participant populations (EBM, ABM, and BGS) and show examples of a raw card sort from a member of each of the three new populations and from members of the two previously described populations. Next, four analyses are presented that provide insights into the differences between how all five populations grouped the cards in their stimulus set, named their constructed groups, and rationalized their card groupings. The four analyses are 1) prevalence of deep, surface, and unexpected card pairings; 2) ED from hypothesized deep-feature and surface-feature sorts; 3) prevalence of hypothesized deep-feature and surface-feature group names; and 4) analysis of card-sorting rationales. For ease of comparison, all of the analyses shown include the two previously published populations, NBM and BF reprinted from Smith *et al.* (2013). The figures, tables, and results are organized to show comparisons among the five participant populations, and comparisons between unframed and framed task conditions for each of the five populations.

### Description of Participant Populations

The participant populations for the current biology card-sort study are described in Table 1. In this study, we collected biology card-sorting data for three new populations: EBM, ABM, and BGS. For ease of comparison with these three new populations, we are including previous data collected from NBM and BF (Smith *et al.*, 2013).

EBM were recruited from Introductory Biology I, the first biology class required for biology majors at the institution. There were 315 EBM who were given the BCST as part of an in-class activity; 31 students did not consent to the use of their data for this study for a participation rate of 90%. Of the remaining 284 students, 99 students were excluded for one or both of the following reasons: 82 students were excluded because they were not biology majors and 32 students were excluded due to sorting anomalies. Sorting anomalies were defined as using a card more than once in either the unframed or framed condition, or not using all of the cards in either the framed or unframed condition. The final population size for EBM was $n = 185$. Of EBM, 70% identified as female and 70% who answered the optional demographic question identified as nonwhite (Table 1).

ABM were recruited from an upper-division genetics course, the last course required for all biology majors. There were 138 ABM who were given the BCST as part of an in-class activity; four students did not consent to the use of their data for this study for a participation rate of 97%. Of the remaining 134 students, 25 were excluded for one or more reasons as follows: 10 students were excluded because they were not biology majors, 15 students' data contained sorting anomalies, one student did not submit a demographics form, and one student did not submit a reflection form for the framed sort. The final population size for ABM was $n = 109$. We are showing data from one ABM class, as the NBM and EBM data were both only collected in a single class. However, two other ABM data sets were collected in the same class but in different years. There were no statistically significant differences between the data sets, but two of the data sets were much more similar to each other than the third data set. Of these two data sets, we are presenting the data set with the greatest number of participants. Of ABM, 59% identified as female and 74% who answered the optional demographic question identified as nonwhite (Table 1).

BGS were recruited from a pool of 165 graduate students in a biology master's degree program. Thirty BGS participated and gave consent for a participation rate of 18%. Of these 30 students, one task contained a sorting anomaly. The final population size for BGS was $n = 29$. Of BGS, 59% identified as female and 32% who answered the optional demographic question identified as nonwhite (Table 1).

Detailed descriptions of the 101 NBM and 23 BF can be found in Smith *et al.*, 2013. There were significant differences in the percentage of female participants among the five populations ($\chi^2 = 20.1$, $df = 4$, $p = 0.0004$). There were also significant differences in the percentage of participants of color between the five populations ($\chi^2 = 25.2$, $df = 4$, $p < 0.0001$).

### Example Card Sorts from a Non–Biology Major, Entering Biology Major, Advanced Biology Major, Biology Graduate Student, and Biology Faculty

Figure 3 shows example BCST responses from a single member of the NBM (Figure 3, A and B), EBM (Figure 3, C and D), ABM (Figure 3, E and F), BGS (Figure 3, G and H), and BF (Figure 3, I and J) groups. NBM (Figure 3, A and B) and BF (Figure 3, I and J) examples are reprinted from Smith *et al.* (2013). The examples in Figure 3 demonstrate the different kinds of information that can be collected from the biology card-sort task. In the unframed condition, participants vary in the number of groups that they sort the cards into, how they name each group, and which cards they put into each group. In the framed condition,

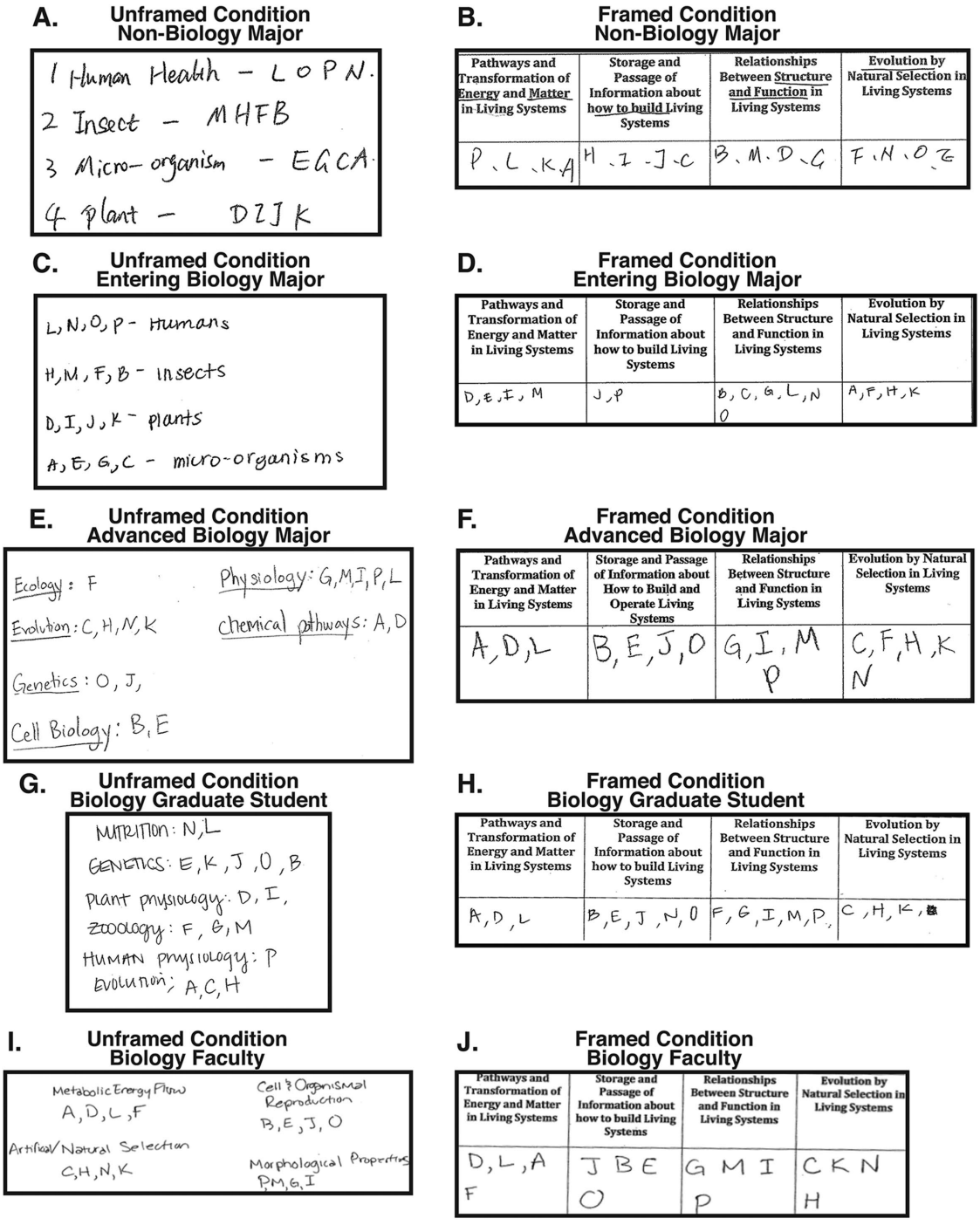


**FIGURE 3. Example BCST results. (A) Unframed condition, non–biology major (NBM). (B) Framed condition, NBM. (C) Unframed condition, entering biology major (EBM). (D) Framed condition, EBM. (E) Unframed condition, advanced biology major (ABM). (F) Framed condition, ABM. (G) Unframed condition, biology graduate student (BGS). (H) Framed condition, BGS. (I) Unframed condition, biology faculty (BF). (J) Framed condition, BF.**

both the number of groups and the name of each group are given to the participant. Figure 3A further shows an example of an exact hypothesized surface-feature card sort, examples of which were found in all undergraduate student populations studied. An exact hypothesized deep-feature card sort was found in the previously published BF population (Figure 3I), and in one of the ABM data collections not presented here.

### Analyses of Prevalence of Surface-Feature, Deep-Feature, and Unexpected Card Pairings

*Percent Card Pairings in the Unframed Card Sort.* In the unframed card-sort condition (Figure 4A and Table 3), EBM ($n$ = 185) generated an average of 37.0 ± 2.1% surface-feature card pairings, 35.7 ± 1.9% deep-feature card pairings, and 27.3 ± 1.1% unexpected card pairings. ABM ($n$ = 109) generated an average

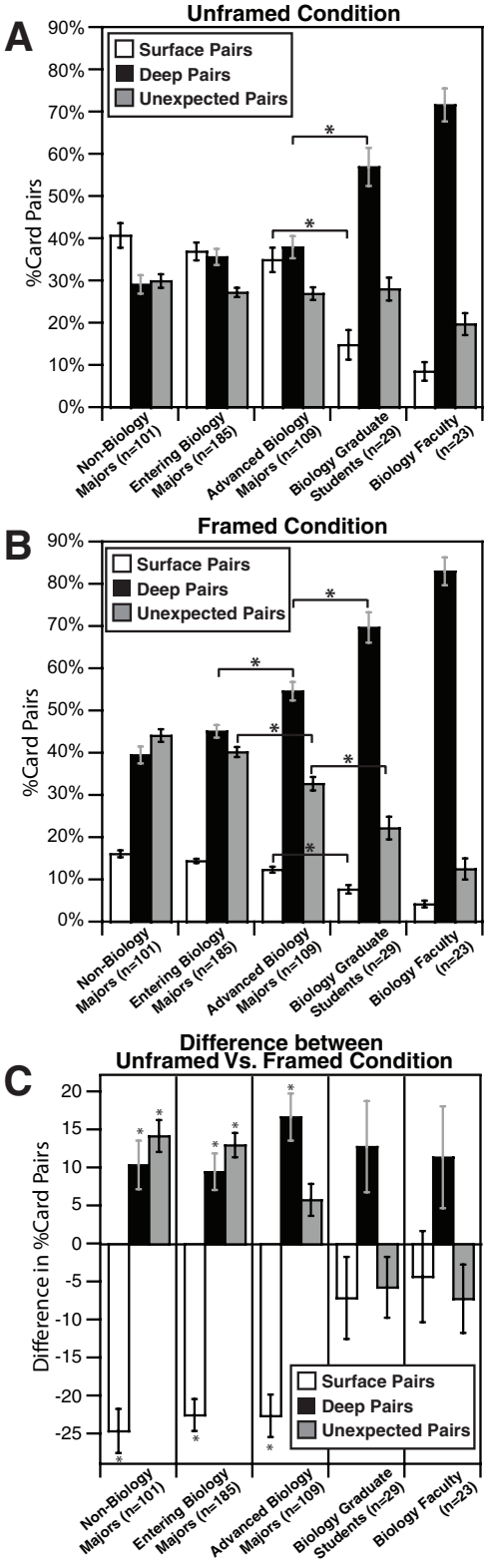


**FIGURE 4. Surface-feature, deep-feature, and unexpected card pairings among populations with increasing biology experience. Mean percentage of surface-feature card pairs (white bars), deep-feature card pairs (black bars), or unexpected card pairs (gray bars) in each population for the unframed condition (A) or framed condition (B). Error bars represent the SEM. Differences between adjacent populations that are significant to a Bonferroni-adjusted $p < 0.004$ (see *Methods*) are marked with an asterisk. (C) Difference between the average percent card pairs in the framed – unframed condition. A positive number indicates an increase in that card pair category after framing. A negative number indicates a decrease in that card pair category after framing. Differences between adjacent populations that are significant to a Bonferroni-adjusted $p < 0.003$ (see *Methods*) are marked with an asterisk.**

of 35.0 ± 2.9% surface-feature card pairings, 38.0 ± 2.6% deep-feature card pairings, and 27.0 ± 1.5% unexpected card pairings. BGS ($n = 29$) generated an average of 14.9 ± 3.5% surface-feature card pairings, 57.0 ± 4.5% deep-feature card pairings, and 28.1 ± 2.7% unexpected card pairings in this unframed sorting condition.

We made statistical comparisons between populations that are the closest in expertise, which we term "adjacent populations": NBM versus EBM, EBM versus ABM, ABM versus BGS, and BGS versus BF. Statistical analysis of differences between adjacent populations in mean percent surface-feature pairs in the unframed condition revealed that, while there was a general downward trend, only the difference between ABM and BGS was statistically significant (Figure 4A, white bars, and Table 3). There were no statistically significant differences in the percentage of deep-feature pairs between any of the undergraduate populations, nor was there a statistically significant difference between BGS and BF (Table 3).

The inverse pattern was found when comparing the mean percent deep-feature pairs between populations in the unframed condition (Figure 4A, black bars, and Table 3). While there was a general upward trend in mean percent surface-feature pairs with increasing formal biology education, only the difference between ABM and BGS was statistically significant (Table 3).

Finally, statistical analysis of unexpected pairs in the unframed condition revealed no significant differences among any of the adjacent populations. (Figure 4A, gray bars, and Table 3).

*Percent Card Pairings in the Framed Card Sort.* To investigate the effect of biological framing on how participants categorize their knowledge, we did the same analysis for the results in the framed condition. In the framed card-sort condition (Figure 4B and Table 3), EBM ($n$ = 185) generated an average of 14.5 ± 0.5% surface-feature card pairings, 45.2 ± 1.5% deep-feature card pairings, and 40.3 ± 1.2% unexpected card pairings. ABM ($n$ = 109) generated an average of 12.4 ± 0.7% surface-feature card pairings, 54.7 ± 2.2% deep-feature card pairings, and 32.8 ± 1.6% unexpected card pairings. BGS ($n$ = 29) generated an average of 7.8 ± 1.0% surface-feature card pairings, 69.8 ± 3.6% deep-feature card pairings, and 22.3 ± 2.7% unexpected card pairings in this framed sorting condition.

Statistical analysis of differences between populations revealed that, even after framing, when analyzing the mean percent surface card pairs in the framed condition, only the difference between ABM and BGS was statistically significant (Figure 4B, white bars, and Table 3). On the other hand, upon framing, the difference in mean percent deep pairs between ABM and EBM and between ABM and BGS were now statistically significant (Figure 4B, black bars, and Table 3). Interestingly, neither the difference between NBM and EBM, nor the

TABLE 3. Prevalence of surface-feature, deep-feature, and unexpected card pairing[a]

| | | Unframed task condition | | | Framed task condition | | |
|---|---|---|---|---|---|---|---|
| Participant type | *n* | Surface | Deep | Unexpected | Surface | Deep | Unexpected |
| Non–biology majors | 101 | 40.8% (2.9) | 29.2% (2.2) | 30.0% (1.6) | 16.2% (0.8) | 39.6% (2.0) | 44.2% (1.5) |
| Entering biology majors | 185 | 37.0% (2.1) | 35.7% (1.9) | 27.3% (1.1) | 14.5% (0.5) | 45.2% (1.5) | 40.3% (1.2) |
| Advanced biology majors | 109 | 35.0% (2.9) | 38.0% (2.6) | 27.0% (1.5) | 12.4% (0.7) | 54.7% (2.2) | 32.8% (1.6) |
| Biology graduate students | 29 | 14.9% (3.5) | 57.0% (4.5) | 28.1% (2.7) | 7.8% (1.0) | 69.8% (3.6) | 22.3% (2.7) |
| Biology faculty | 23 | 8.6% (2.2) | 71.7% (3.9) | 19.8% (2.6) | 4.3% (0.8) | 83.1% (3.3) | 12.6% (2.5) |

[a]Adjacent populations with percent card-pair differences that were significant to a Bonferroni-adjusted $p < 0.0042$ (see *Methods*) are denoted with a vertical line and a number. The numbers correspond to the following $p$ values: [1]$p = 0.0007$; [2]$p = 0.0003$, [3]$p = 0.0017$, [4]$p = 0.0002$, [5]$p = 0.0005$, [6]$p < 0.0001$, [7]$p = 0.0014$.

difference between BGS and BF was statistically significant (Table 3). Finally, when analyzing the mean percent unexpected card pairs, the differences between EBM and ABM and between ABM and BGS were statistically significant (Figure 4B, gray bars, and Table 3).

*Comparison of Percent Card Pairings between the Unframed and Framed Card Sorts.* To determine how framing affected the proportion of surface, deep, or unexpected pairs generated by the populations, we calculated the difference between the percent card pairs in the framed versus the unframed condition as shown in Figure 4C. A negative difference indicates that there were more card pairs of that type in the unframed condition compared with the framed condition. A positive difference indicates that there were more card pairs of the type in the framed condition compared with the unframed condition. An asterisk denotes that the difference in percent card pairs between the unframed and framed condition is statistically significant to a Bonferroni-adjusted $p < 0.003$ (see *Methods*). This representation shows how card pairs are redistributed in the framed sort compared with the unframed sort. For example, in NBM, the difference in percent surface pairs between the unframed and framed condition was –24.6 ± 2.9%, the difference in percent deep pairs was 10.4 ± 3.2%, and the difference in percent unexpected pairs was 14.2 ± 2.1%. This indicates that, when shifting to the framed condition, NBM move away from surface pairs and toward deep and unexpected pairs approximately equally. Similarly, in EBM, the difference in percent surface pairs between the unframed and framed condition was –22.5 ± 2.1%, the difference in percent deep pairs was 9.5 ± 2.4%, and the difference in percent unexpected pairs was 13.0 ± 1.6%. All of the differences in percent card pairs between the unframed and framed conditions for NBM and EBM were statistically significant to a Bonferroni-adjusted $p < 0.003$ (see *Methods*; Figure 4C).

Similarly, in ABM, the difference in percent surface pairs between the unframed and framed condition was –22.6 ± 2.8%. However, in ABM, the difference in percent deep pairs was 16.7 ± 3.1%, and the difference in percent unexpected pairs was 5.8 ± 2.1%. The difference in percent surface pairs and percent deep pairs was statistically significant to a Bonferroni-adjusted $p < 0.003$ (see *Methods*), but the difference in percent unexpected pairs was not statistically significant. These results suggest that, in ABM, there is a shift from surface pairs to deep pairs more often than to unexpected pairs.

Neither BGS nor BF showed statistically significant differences in their percent card pairs in the unframed and framed conditions.

### Analyses of EDs from the ED-Surface and the ED-Deep Sorts

An alternate method we used to analyze participants' card groupings was to measure ED from the exact hypothesized surface-feature sort (ED-Surface) and an ED from the exact hypothesized deep-feature sort (ED-Deep) for each participant. As described in *Methods*, the ED from an exact hypothesized sort is the minimum number of card moves necessary to turn an individual's card sort into the hypothesized sort. For example, a participant with an exact hypothesized deep-feature sort would have an ED-Deep of 0 necessary card moves, and an ED-Surface of 12 necessary card moves. From these calculated EDs, an average ED-Surface and ED-Deep was calculated and compared for each of the populations, as well as between the unframed and framed task conditions for each population. Units of ED are "necessary card moves."

*ED in the Unframed Card Sort.* EBM constructed card sorts with an average ED-Surface of 7.4 ± 0.2 and an average ED-Deep of 7.4 ± 0.2 (Figure 5A). ABM constructed card sorts with an average ED-Surface of 7.6 ± 0.3 and an average ED-Deep of 7.4 ± 0.3. BGS constructed card sorts with an average ED-Surface of 10.0 ± 0.4 and an average ED-Deep of 5.4 ± 0.5.

Statistical comparison of these means showed that, in the unframed condition, for both ED-Deep and ED-Surface, only the difference between ABM and BGS was statistically significant. The ED-Deep and ED-Surface of the three undergraduate populations (NBM, EBM, and ABM) were statistically indistinguishable (Table 4). The unframed ED-Deep and ED-Surface of BGS and BF were also statistically indistinguishable from each other (Table 4).

*ED in the Framed Card Sort.* To investigate the effect of biological framing on how participants group their cards, we again calculated the ED-Surface and ED-Deep for participant card groupings in the framed condition. Here, we report that, in the framed condition, EBM constructed card sorts with an average ED-Surface of 9.3 ± 0.1 and an average ED-Deep of 5.3 ± 0.2 (Figure 5B). ABM constructed card sorts with an average ED-Surface of 9.7 ± 0.1 and an average ED-Deep of 4.0 ± 0.2. BGS constructed card sorts with an average ED-Surface of 10.3 ± 0.2 and an average ED-Deep of 2.4 ± 0.3.

Statistical comparison of these means showed that, upon framing, the ED-Deep of ABM became statistically significantly lower than that of EBM (Table 4). This indicates that, upon framing, ABM sorts are closer to a perfect hypothesized expert sort compared with EBM.

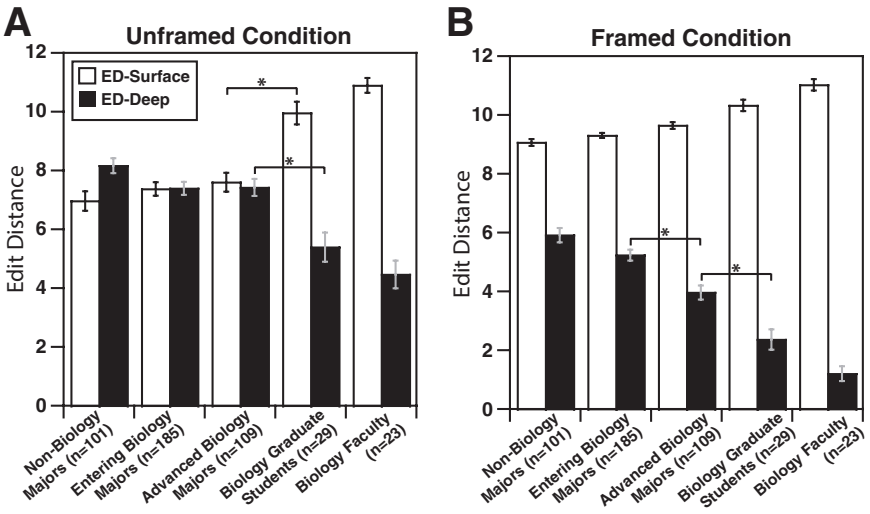


**FIGURE 5. Edit distance to the perfect hypothesized expert sort (ED-Deep) or perfect hypothesized novice sort (ED-Surface). ED-Deep (black bars) and ED-Surface (white bars) in the (A) unframed and (B) framed conditions. Units of ED are "necessary card moves." Error bars represent the SEM. Differences between adjacent populations that are significant to a Bonferroni-adjusted $p < 0.006$ (see *Methods*) are marked with an asterisk.**

ED-Surface in the framed conditions showed very few statistically significant differences. Only the differences between ABM and BGS were statistically significant (Figure 5B and Table 4).

*Comparison of ED between the Unframed and Framed Card Sorts.* To determine how framing affected the ability of populations to sort in a more novice or more expert-like manner, we calculated the difference in EDs between the unframed and framed conditions. A negative difference indicates that framing caused the population to move away from the perfect hypothesized sort. A positive difference indicates that framing led the population to move toward the perfect hypothesized sort (Table 4).

Comparison of ED-Surface and ED-Deep analyses for the unframed and framed task conditions revealed significant shifts between the two task conditions for all undergraduate populations (Table 4). NBM, EBM, and ABM all shifted toward perfect deep-feature sorts (smaller ED-Deep) and away from perfect surface-feature sorts (bigger ED-Surface). In all three populations, the difference between the unframed and framed conditions in both ED-Surface and ED-Deep was statistically significant to a Bonferroni-adjusted $p < 0.005$ (see *Methods*).

For BGS and BF, their shifts away from a perfect surface-feature sort upon framing were not statistically significant. However, framing allowed both of these populations to move toward perfect deep-feature sorts, and for both populations, the difference in ED-Deep between the unframed and framed conditions was statistically significant to a Bonferroni-adjusted $p < 0.005$ (see *Methods*; Table 4).

### Analyses and Comparison of Constructed Card-Group Names

The metrics described above give insights into how participants grouped their cards in both the unframed and framed conditions. Another source of information is how participants chose to name their groups in the unframed condition. Quantifying the frequency with which specific hypothesized surface or deep features are used as group names can give insights into how different populations categorize their biological knowledge. Below we quantify the prevalence of the four hypothesized deep features and the prevalence of the four hypothesized surface features in the group names that participants assigned their card groups in the unframed condition.

*Hypothesized Deep-Feature Group Names.* Analysis of the prevalence of card-group names related to the four hypothesized deep features (see columns in Figure 1) is shown in Figure 6. With the exception of the deep feature "storage and passage of information," the use of each deep-feature card category was less than 50% among every student population.

For the deep feature "evolution and natural selection," the prevalence of this group name among the NBM (20.8%), EBM (22.7%), and ABM (22.0%) was statistically indistinguishable ($\chi^2 = 0.14$, $df = 2$, $p = 0.293$; Figure 6A). Compared with the undergraduate populations, the group name "evolution and natural selection" was twice as prevalent among BGS (48.3%) and four times as prevalent among BF

**TABLE 4. Edit distances (EDs) from the hypothesized surface-feature sort and the hypothesized deep-feature sort[a]**

| | | ED from surface sort | | | ED from deep sort | | |
|---|---|---|---|---|---|---|---|
| **Participant type** | ***n*** | **Unframed** | **Framed** | **Difference[b]** | **Unframed** | **Framed** | **Difference[b]** |
| Non–biology major | 101 | 7.0 (0.3) | 9.1 (0.1) | −2.1* (0.3) | 8.2 (0.3) | 5.9 (0.2) | 2.2* (0.4) |
| Entering biology major | 185 | 7.4 (0.2) | 9.3 (0.1) | −1.9* (0.2) | 7.4 (0.2) | 5.3 (0.2) | 2.2* (0.3) |
| Advanced biology major | 109 | 7.6 (0.3) | 9.7 (0.1) | −2.0* (0.2) | 7.4 (0.3) | 4.0 (0.2) | 3.5* (0.4) |
| Biology graduate students | 29 | 10.0 (0.4) | 10.3 (0.2) | −0.4 (0.6) | 5.4 (0.5) | 2.4 (0.3) | 3.0* (0.7) |
| Biology faculty | 23 | 10.9 (0.3) | 11.0 (0.2) | −0.1 (0.7) | 4.5 (0.5) | 1.2 (0.3) | 3.3* (0.8) |

[a]Note that lower ED numbers indicate sorts more similar to the hypothesized sort. Differences between the unframed and framed conditions that were significant to a Bonferroni-adjusted $p < 0.005$ (see *Methods*) are denoted with an asterisk. NBM and BF data are reprinted from Smith *et al.* (2013). Adjacent populations with ED differences that were significant to a Bonferroni-adjusted $p < 0.006$ (see *Methods*) are denoted with a vertical line and a number. The numbers correspond to the following $p$ values: [1]$p = 0.0004$; [2]$p = 0.0049$; [3]$p = 0.0007$; [4]$p < 0.0001$; [5]$p = 0.0016$.

[b]Difference represents the difference in ED between the unframed and framed conditions. A negative number denotes that the population moved away from the hypothesized perfect sort in the framed condition. A positive number denotes that the population moved toward the hypothesized perfect sort in the framed condition.

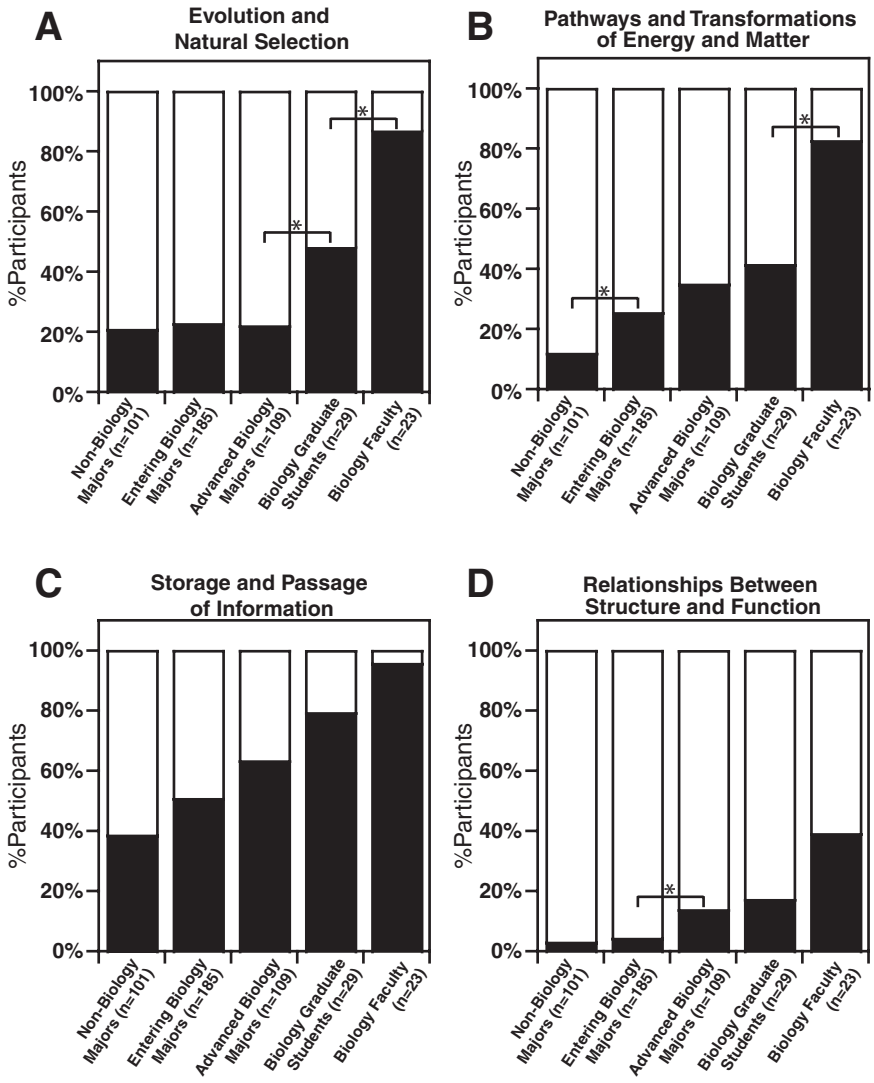


**FIGURE 6. Prevalence of deep-feature card-group names in the unframed condition. In the unframed sort, the proportion of participants who did (black bars) or did not (white bars) include a group name corresponding to one of the four hypothesized deep features: evolution and natural selection (A), pathways and transformations of energy and matter (B), storage and passage of information (C), and relationships between structure and function (D). For each panel, differences between adjacent populations that are significant to a Bonferroni-adjusted $p < 0.0125$ (see *Methods*) are marked with an asterisk.**

"Storage and passage of information" was by far the most prevalent group name. We found that 38.6% of NBM, 50.8% of EBM, 63.3% of ABM 79.3% of BGS, and 95.7% of BF used a group name that fell into this category (Figure 6C). The differences among all populations were statistically significant ($\chi^2 = 37.6$, $df = 4$, $p < 0.0001$). However, none of the differences between adjacent populations were statistically significant. It should be noted that the group name "genetics" was accepted in this category (Supplemental Table S1), and that more than 90% of respondents in all populations who were coded as having a "storage and passage of information" group name used the category name "genetics" (unpublished data).

Finally, "relationships between structure and function" was by far the least prevalent group name, with only 3.0% of NBM, 4.3% of EBM, 13.8% of ABM, 17.2% of BGS, and 39.1% of BF using a group name that fell into this category (Figure 6D). Only the difference between EBM and ABM was statistically significant ($\chi^2 = 9.51$, $df = 1$, $p = 0.002$).

*Hypothesized Surface-Feature Group Names.* Analysis of the prevalence of card-group names related to the four hypothesized surface features (see rows in Figure 1) is shown in Figure 7. Surface-feature group names appeared in a similar proportion in NBM, EBM, and ABM, and these populations were statistically indistinguishable from one another by these metrics. Surface-feature group names appeared in a much smaller proportion of BGS and BF, and these two populations were indistinguishable from each other by these metrics.

Specifically, the surface feature "human" appeared in the group names of similar proportions of NBM (47.5%), EBM (37.3%), and ABM (42.2%, $\chi^2 = 2.87$, $df = 2$, $p = 0.24$; Figure 7A). The surface feature "human" also appeared in a similar proportion of BGS (10.3%) and BF (8.7%, $\chi^2 = 0.04$, $df = 1$, $p = 0.84$). Of the adjacent populations, only the difference between ABM and BGS was statistically significant ($\chi^2 = 8.11$, $df = 1$, $p = 0.0044$).

Similarly the surface feature "insect" appeared in the group names of similar proportions of NBM (43.6%), EBM (38.9%), and ABM (32.1%, $\chi^2 = 3.0$, $df = 2$, $p = 0.23$; Figure 7B). The surface feature "insect" also appeared in a similar proportion of BGS (6.9%) and BF (0%, $\chi^2 = 1.6$, $df = 1$, $p = 0.19$). Again, of the adjacent populations, only the difference between ABM and BGS was statistically significant ($\chi^2 = 12.2$, $df = 1$, $p = 0.0005$).

(87.0%). Only the difference between ABM and BGS ($\chi^2 = 7.27$, $df = 1$, $p = 0.007$) and the difference between BGS and BF ($\chi^2 = 7.24$, $df = 1$, $p = 0.007$) were statistically significant (Figure 6A).

In the case of the deep feature "pathways and transformations of energy and matter," we observed a gradual increase in the prevalence of this group name with increasing biology education: NBM (11.9%), EBM (25.4%), ABM (34.9%), and BGS (41.4%; Figure 6B). However, in BF, the prevalence of a "pathways and transformation of energy and matter" group name was 82.6%, or about twice that of BGS (41.4%). Only the difference between NBM and EBM ($\chi^2 = 8.48$, $df = 1$, $p = 0.0036$) and the difference between BGS and BF ($\chi^2 = 9.06$, $df = 1$, $p = 0.0026$) were statistically significant.

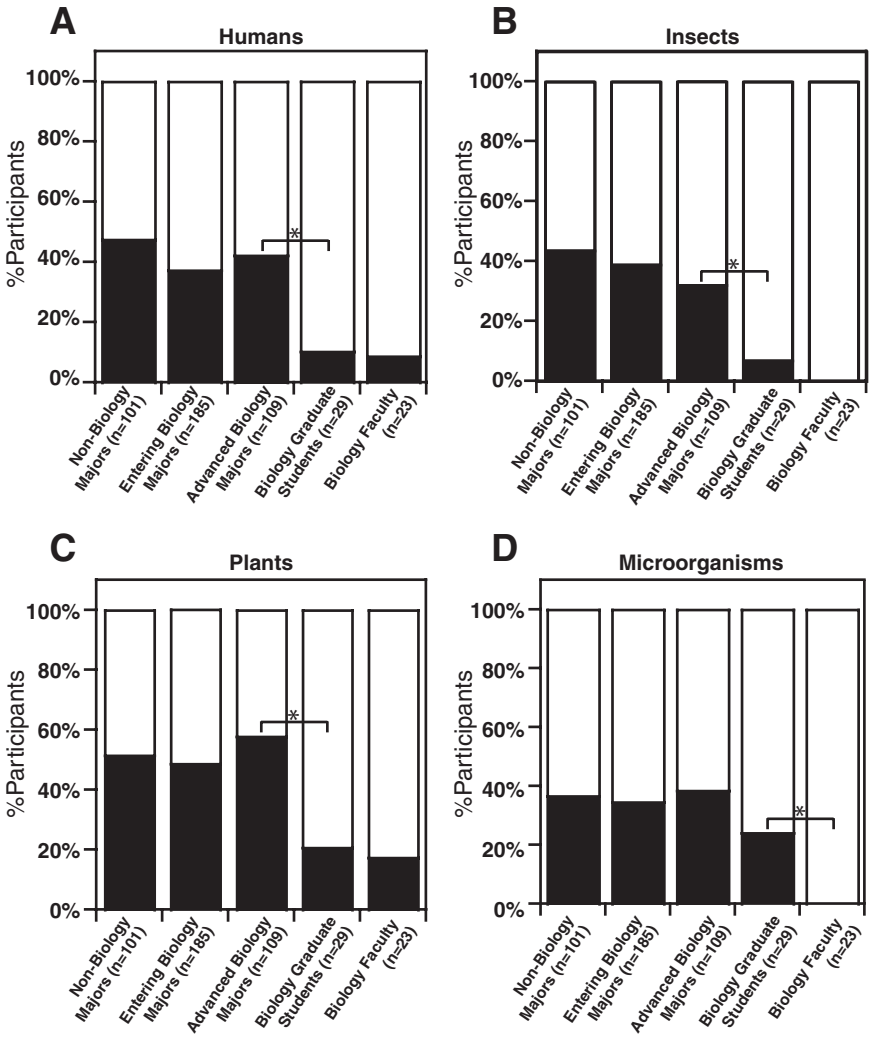


**FIGURE 7. Prevalence of surface-feature card-group names in the unframed condition. In the unframed sort, the proportion of participants who did (black bars) or did not (white bars) include a group name corresponding to one of the four hypothesized surface features: humans (A), insects (B), plants (C), and microorganisms (D). For each panel, differences between adjacent populations that are significant to a Bonferroni-adjusted $p < 0.0125$ (see *Methods*) are marked with an asterisk.**

The surface feature "plants" appeared in the group names of similar proportions of NBM (47.5%), EBM (37.3%), and ABM (42.2%, $\chi^2 = 2.9$, $df = 2$, $p = 0.24$; Figure 7C). The surface feature "plants" also appeared in a similar proportion of BGS (20.7%) and BF (17.4%, $\chi^2 = 10.1$, $df = 1$, $p = 0.76$). Again, of the adjacent populations, only the difference between ABM and BGS was statistically significant ($\chi^2 = 8.18$, $df = 1$, $p = 0.0042$).

Finally, the surface feature "microorganisms" appeared in the group names of similar proportions of NBM (36.6%), EBM (34.6%), ABM (38.5%), and BGS (24.1%, $\chi^2 = 2.2$, $df = 3$, $p = 0.53$; Figure 7D), while 0% of BF had a group named "microorganisms." For this group name, only the difference between BGS and BF was statistically significant ($\chi^2 = 11.1$, $df = 1$, $p = 0.0009$).

### Analyses and Comparison of Card-Sorting Strategy Explanations from Responses to Posttask Reflection Questions

To gain a deeper understanding of how participants organize their biological knowledge, participants were asked in a posttask reflection question: 1) Describe why you grouped certain problems together. Give an example of your reasoning; and 2) How did you decide on the names of your groups? As detailed in *Methods*, the answers were analyzed and coded using a grounded theory approach. Using this approach, we found that, in addition to providing rationales based on the hypothesized surface features and hypothesized deep features, sizable fractions of certain populations also used an explicit curricular rationale (sorting based on university course titles, textbook chapter titles, etc.; Table 2).

*Hypothesized Surface-Feature Rationales.* For the surface feature–based rationales, we found that four times the proportion of NBM (37.6%) used surface-feature rationales compared with BF (8.7%; Table 2). Still, there were no significant differences between the use of hypothesized surface-feature rationales by adjacent populations, with 33% of EBM, 32.1% of ABM, and 13.8% of BGS also using these types of rationales (Table 2).

*Hypothesized Deep-Feature Rationales.* For hypothesized deep-feature rationales, we found that 22.8% of NBM, 37.8% of EBM, 32.1% of ABM, 48.3% of BGS, and 100% of BF used such rationales (Table 2). The differences between NBM and EBM ($\chi^2 = 6.8$, $df = 1$, $p = 0.0093$) and between BGS and BF ($\chi^2 = 16.7$, $df = 1$, $p < 0.0001$) were found to be statistically significant (Table 2).

*Unhypothesized Rationale: Curricular Rationales.* There were also many differences between populations in the prevalence of curricular rationales in the card sorting–strategy explanations. This category included sorting strategies based on undergraduate or graduate course titles, textbook chapter titles, and so on. Unsurprisingly, the prevalence of this category among students increased with increasing formal biology education: 6.9% of NBM, 10.3% of EBM, 28.4% of ABM, and 48.3% of BGS reported using curricular rationales in the sorting of their cards, as did 39.1% of BF (Table 2). Interestingly, the only statistically significant difference between adjacent populations was between EBM and ABM ($\chi^2 = 16.0$, $df = 1$, $p < 0.0001$; Table 2).

## DISCUSSION

Instructors and researchers have many choices for assessing student understanding of biological concepts, including constructed-response questions and concept inventories (Smith and Tanner, 2010). Unlike many methods of assessment however, the BCST probes the frameworks that students use to organize their conceptual biological knowledge, rather than the content of that knowledge itself. Previous research has demonstrated the ability of the BCST to differentiate between NBM and BF, two populations hypothesized to be on the extreme ends of the scale of biology expertise among adults (Smith *et al.*, 2013). Here, we used the BCST to investigate whether we could detect differences in conceptual biology expertise among students at intermediate points during their biology education. To this end, we administered the BCST to three additional populations: EBM, ABM, and BGS. In the following sections, we describe new insights into how students organize their conceptual biological knowledge gleaned from these additional populations and discuss the implications of our findings.

### ABM Demonstrated More Deep-Feature Card Sorts Than EBM, but Only When Provided with Deep-Feature Categories

How does undergraduate biology education influence the way ABM organize their biological conceptual knowledge compared with EBM? One might assume that, as a result of their undergraduate course work, ABM would organize their biological conceptual knowledge more like putative biology experts. However, our results indicate that an undergraduate biology education does not seem to lead to a shift in the way biology students think toward a more expert-like or deep-feature framework when organizing their knowledge without guidance. This is a somewhat troubling but not necessarily surprising finding. Several studies have shown that biology students improve in their content knowledge and acquire more expert-like attitudes toward science (as measured by the CLASS-Bio attitudinal survey) during their undergraduate biology education (Marbach-Ad *et al.*, 2010; Hansen and Birol, 2014; Newman *et al.*, 2016). However, this is not always the case (Garvin-Doxas and Klymkowsky, 2008; Abraham *et al.*, 2014). Furthermore, a prominent study using the Collegiate Learning Assessment—a standardized test designed to measure general (as opposed to discipline-specific) critical thinking—showed that students' scores on this exam showed no statistically significant increase during the first 2 years of their college education (Arum and Roksa, 2011). Taken together, these results suggest that, while students may be gaining content knowledge as a result of their undergraduate education, it is unclear whether they are organizing that content knowledge in ways that disciplinary experts might expect. In fact, to what extent do instructors regularly teach students how to connect ideas "among seemingly disparate pieces of information, concepts, and questions," as *Vision and Chang*e (AAAS, 2011) recommends?

Intriguingly, while NBM, EBM, and ABM were statistically indistinguishable by all of the metrics analyzed in the unframed condition, once the subjects were given the hypothesized deep feature–based categories and asked to sort the cards into these four categories, ABM were statistically distinguishable from EBM and NBM. Specifically, ABM generated more deep-feature pairs and had smaller EDs to the hypothesized deep-feature sort compared with EBM. These results suggest that ABM are perhaps able to use the hypothesized framework if it is provided to them, but that they have not developed and do not use this framework themselves unprompted. Alternatively, ABM may have developed this framework, but they do not readily recall it without cuing. A major question that arises from our observation that ABM—but not EBM—are able to use the deep-feature framework once it is presented to them, is whether this is evidence of student learning, or whether this represents a process of student selection during education. Are ABM able to use the deep-feature framework because of what they have learned during their time in college? Or are students who were able to use the framework in the first place somehow "selected" for in undergraduate environments, enabling them to persist as biology majors? Our results further raise the question: If ABM are not using the deep-feature framework to organize their conceptual biology knowledge, what kinds of frameworks are they using, and how do those frameworks differ—if at all—from the frameworks used by other student populations?

### The Largest Differences in Conceptual Expertise Were between ABM and BGS

Our results demonstrate that BGS organize their biology knowledge using a deep-feature framework to a much greater extent than do any of the undergraduate populations. Specifically, we showed that in the absence of external prompting, the only adjacent populations that were statistically distinguishable in any of the quantitative analytical metrics of the card sorts were ABM and BGS. So, if there are few differences between undergraduate populations in how students organize their biological knowledge in the absence of external prompting, but we know that NBM and BF organize their knowledge very differently, when does this reorganization happen? Our results suggest either that students undergo significant reorganization of their conceptual biology knowledge sometime between being ABM and BGS, or that students who choose to go on to graduate school are more likely to have developed a deep-feature conceptual framework. (Or, at an extreme, that these students already possessed this framework when entering college.) We propose that this phenomenon of BGS and BF having much more similar ways of organizing their biology knowledge than ABM and BF could be an example of "self-selection," wherein students who have organizational frameworks that more closely resemble those of faculty are more likely to have a science identity and the self-efficacy required to choose to pursue graduate education in biology (Tanner and Allen, 2004; Trujillo and Tanner, 2014).

Interestingly, we found no statistically significant differences in the quantitative analytical metrics of the card sorts between BGS and BF, suggesting that both populations use the deep-feature conceptual framework to a similar extent, consistent with the idea that acquiring a deep-feature knowledge framework is a key part of continuing on to postgraduate studies and a career in academic science.

### There Are Few Differences in How NBM and EBM Organize Their Conceptual Biology Knowledge

Another notable finding was that there were no apparent differences in how NBM and EBM organized their biological knowledge. Specifically, in both the unframed and framed conditions, there were no statistically significant differences between NBM

and EBM in the percent surface, deep, or unexpected card pairings, nor was there a difference between these two populations in ED-Surface or ED-Deep. These results suggest that NBM and EBM do not organize their biology knowledge in fundamentally different ways, an important finding that contributes to the body of research on whether and how biology majors and nonmajors are different from one another as they enter higher education (Sundberg and Dini, 1993; Knight and Smith, 2010). Our results suggest that NBM and EBM are indistinguishable in terms of their biological knowledge framework, and therefore our results provide an important reminder that our biology majors may not be as advanced as we might think compared with our nonmajors students, and that distinctions we may make between these groups may not be functionally important.

### The BCST Revealed a New, Unhypothesized Organizational Framework: Curricular Content

One advantage of the BCST over other assessment approaches is that the BCST is able to reveal unhypothesized knowledge frameworks that participants use to organize their biological knowledge. This is because participants are asked to provide a rationale for why they grouped certain cards together after they sorted the 16 cards. We originally hypothesized that participants with greater biological experience would describe a sorting rationale based on hypothesized deep features, while participants with less biological experience would describe a sorting rationale based on surface features. This was true for BF, 100% of whom described using a sorting strategy based on deep features (Table 2). However, undergraduate students used both surface-feature rationales and deep-feature rationales, but less than 40% of the time.

So, what other sorting rationales do students report? Rather than the hypothesized surface-feature rationales, we found that students used a variety of other rationales when describing their card-sorting strategies. The most prominent framework that was brought up by students was a curricular framework. Individuals were coded as having a curricular framework if they reported that they sorted their cards based on explicit curricular rationales, including in what class they would expect to see the questions, or in which chapter of a textbook they would expect to find the problem.

Interestingly, among the three rationales that we coded for, the curricular rationale was the only one in which there was a statistically significant shift between EBM and ABM. Certainly this makes intuitive sense, as in the transition between high school and college, classes go from being called "biology" to having more specific names, and from a student's perspective, it would stand to reason that departments and publishers would structure their classes and textbooks around important divisions in biology. As reform efforts in undergraduate biology education gain more momentum, our results suggest that careful choice of the organizational frameworks of classes—such as organizing and naming classes based on the organizing principles laid out in *Vision and Change* (AAAS, 2011)—could be an important contribution to helping students develop a deep feature–based framework for organizing their conceptual biology knowledge.

### Implications

*Vision and Change* (AAAS, 2011) was a collaboratively produced document that proposed five big-picture organizing principles in biology. The BCST can be used to measure to what extent students use four of those core concepts to organize their biological knowledge. The finding that, in the absence of cuing, the conceptual frameworks of ABM do not seem to be based on deep features to any greater extent than those of EBM is troubling and suggests opportunities for improvement. After all, when students go out into the real world and need to make decisions about genetically modified organisms, personalized medicine, climate change, or any number of other real-world issues that we might think their biology education should prepare them to address, they will not have anyone there to provide the big-picture framework for them. What might we be able to do as instructors to ensure that our students are able to employ that deep-feature framework for themselves? Many of our ABM and BGS used an explicit curricular framework to organize their biological conceptual knowledge, but to what extent do our course names or textbooks align with the organizing principles set forth in *Vision and Change* (Wagner *et al.*, 2015)? To what extent do we explicitly organize the material we teach within the *Vision and Change* framework? Or are our course names and textbooks organized around surface features, such as "plant biology" or "human biology"? And to what extent do we give our students practice connecting the content in our classrooms to the *Vision and Change* framework?

### Limitations and Future Directions

One limitation of this study and the previous BCST study is that they were both conducted in the context of a single, large, diverse, master's-granting institution where more than 89% of biology department instructors have undergone at least 40 h of pedagogical training (unpublished data). This raises a number of interesting questions. For example, how might faculty at other institutions, where there are fewer teaching responsibilities, organize their biology knowledge differently from the faculty experts in Smith *et al.* (2013)? And how might the students at those institutions organize their knowledge differently from the students in these studies? Another limitation of conducting this study at a single institution is that we could not measure the effect of different curricula on the frameworks that students use to organize their biology knowledge. It is tempting to speculate that the organization of the content of the biology courses that students take might affect how students organize their biology knowledge. But are students who learn biology in a department that has adopted a curriculum more in line with the suggestions in *Vision and Change* (AAAS, 2011) more likely to organize their biology knowledge like experts compared with students who are taught using a more traditional curriculum? A final important variable is pedagogical differences between instructors. Many studies have demonstrated that classes that take an active-learning approach result in improved student outcomes compared with traditional lecture-only classes (Freeman *et al.*, 2014). How might students in classrooms where their instructor has been trained in effective pedagogy perform differently on the BCST specifically, compared with students taught in a more traditional setting?

Another limitation of this work is that it is a cross-sectional study, and therefore we cannot make specific conclusions about how individual students may or may not be changing how they organize their conceptual biology knowledge over time. As mentioned previously, a major question that arises from our

observation that ABM—but not EBM—are able to use the deep-feature framework once it is presented to them, is whether this is evidence of student learning, or whether this represents a process of student selection during education. One way to investigate these questions would be to conduct a longitudinal study in which the BCST is administered to a cohort of EBM and then administered again when these same students are ABM. If the same students who were unable to use the framework as EBM are able to use the framework as ABM, this would be evidence for the idea that students are learning. On the other hand, if the students who are able to use the framework as ABM are primarily those same students who could use the framework as EBM, this may indicate that those students were selected for and persisted. This work is currently ongoing in our labs.

### Conclusions

We have demonstrated the utility of the BCST in assessing conceptual expertise in biology among three distinct student populations: EBM, ABM, and BGS. Our results indicate that the BCST will be a powerful class or departmental assessment tool, enabling individual instructors or entire departments to answer the question of whether students are becoming more expert-like in their thinking as a result of a single biology course or as a result of their entire biology education. In fact, the BCST has already begun to be used in class-based assessments, which have demonstrated that the BCST is able to detect changes in students' conceptual biology expertise over the course of a single semester (Hoskinson, personal communication). In the future, longitudinal studies using the BCST will enable departments to assess whether students are changing the way they organize their knowledge as a result of their biology education, or whether we are simply selecting for students who already have a more expert-like organizational framework, and failing to retain students who do not develop that framework. Answering this question will be a critical contribution toward understanding how—or if— biology education affects how students organize their conceptual biology knowledge.

## ACKNOWLEDGMENTS

This work was supported by National Science Foundation CAREER Award #DRL-0954127. We thank Dr. Seth Bush and Dr. Gregory Scott for helpful comments on the paper. We also thank Science Education Partnership and Assessment Laboratory (SEPAL) staff and students for their comments on this work. Finally, we thank all of the faculty and students who participated in this research.

## REFERENCES

Abraham JK, Perez KE, Price RM (2014). The Dominance Concept Inventory: a tool for assessing undergraduate student alternative conceptions about dominance in Mendelian and population genetics. CBE Life Sci Educ 13, 349–358.

Ambrose SA, Bridges MW, DiPietro M, Lovett MC, Norman MK (2010). How does the way students organize knowledge affect their learning? How Learning Works: Seven Research-Based Principles for Smart Teaching, San Francisco, CA: Jossey-Bass, 40–65.

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

Arum R, Roksa J (2011). Academically Adrift: Limited Learning on College Campuses, Chicago: University of Chicago Press.

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

Clapper B (2008). munkres: munkres algorithm for the assignment problem, version 1.0.5.4. http://software.clapper.org/munkres (accessed 12 February 2017).

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

Eylon B, Reif F (1984). Effects of knowledge organization on task performance. Cogn Instr 1, 5–44.

Freeman S, Eddy SL, McDonough M, Smith MK, Okoroafor N, Jordt H, Wenderoth MP (2014). Active learning increases student performance in science, engineering, and mathematics. Proc Natl Acad Sci USA 111, 8410–8415.

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.

Hansen MJ, Birol G (2014). Longitudinal study of student attitudes in a biology program. CBE Life Sci Educ 13, 331–337.

Hardiman PT, Dufresne R, Mestre JP (1989). The relation between problem categorization and problem solving among experts and novices. Mem Cognit 17, 627–638.

Irby SM, Phu AL, Borda EJ, Haskell TR, Steed N, Meyer Z (2016). Use of a card sort task to assess students' ability to coordinate three levels of representation in chemistry. Educ Chem 17, 337–352.

Knight JK, Smith MK (2010). Different but equal? How nonmajors and majors approach and learn genetics. CBE Life Sci Educ 9, 34–44.

Krieter FE, Julius RW, Tanner KD, Bush SD, Scott GE (2016). Thinking like a chemist: development of a chemistry card-sorting task to probe conceptual expertise. J Chem Educ 93, 811–820.

Kuhn HW (2010). The Hungarian method for the assignment problem. In: 50 Years of Integer Programming 1958–2008: From the Early Years to the State-of-the-Art, New York: Springer, 29–47.

Marbach-Ad G, McAdams KC, Benson S, Briken V, Cathcart L, Chase M, El-Sayed NM, Frauwirth K, Fredericksen B, Joseph SW, *et al.* (2010). A model for using a concept inventory as a tool for students' assessment and faculty professional development. CBE Life Sci Educ 9, 408–416.

Mason A, Singh C (2011). Assessing expertise in introductory physics using categorization task. Phys Rev Spec Top Phys Educ Res 7, 1–17.

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.

Python Software Foundation (2011). Python, version 2.7.2. www.python.org (accessed 14 October 2013).

SAS Institute (2015). JMP 12, version 12.1.0. Cary, NC.

Smith JI, Combs ED, Nagami PH, Alto VM, Goh HG, Gourdet Maa, 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 MU (1990). Knowledge structures and the nature of expertise in classical genetics. Cogn Instr 7, 287–302.

Sundberg MD, Dini ML (1993). Science majors vs nonmajors: is there a difference? J Coll Sci Teach 23, 299–304.

Tanner K, Allen D (2004). Approaches to biology teaching and learning: earning styles and the problem of instructional selection—engaging all students in science courses. Cell Biol Educ 3, 197–201.

Trujillo G, Tanner KD (2014). Considering the role of affect in learning: monitoring students' self-efficacy, sense of belonging, and science identity. CBE Life Sci Educ 13, 6–15.

Wagner JD, Campbell AM, Sly BJ, Paradise CJ (2015). An active textbook converts "vision and tweak" to vision and change. CourseSource 2, 1–6.

Weiser M, Shertz J (1983). Programming problem representation in novice and expert programmers. Int J Man Mach Stud 19, 391–398.