CBE—Life Sciences Education
Vol. 12, 628–644, Winter 2013

Article

Development of the Biology Card Sorting Task to Measure Conceptual Expertise in Biology

Julia I. Smith,* Elijah D. Combs,† Paul H. Nagami,† Valerie M. Alto,* Henry G. Goh,* Muryam A. A. Gourdet,* Christina M. Hough,* Ashley E. Nickell,* Adrian G. Peer,* John D. Coley,‡ and Kimberly D. Tanner†

*Holy Names University, Oakland, CA 94619; †San Francisco State University, San Francisco, CA 94132; ‡Northeastern University, Boston, MA 02115

Submitted June 5, 2013; Revised July 28, 2013; Accepted July 29, 2013
Monitoring Editor: A. Malcolm Campbell

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

INTRODUCTION

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

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

In terms of the conceptual domains referred to in this excerpt, *Vision and Change* goes on to articulate and specify that all undergraduates should develop a basic conceptual understanding of the following core biological concepts:

DOI: 10.1187/cbe.13-05-0096
Address correspondence to: Kimberly D. Tanner (kdtanner@sfsu.edu).

1) evolution, 2) structure and function, 3) information flow, exchange, and storage, 4) pathways and transformations of energy and matter, and 5) systems (AAAS, 2011).

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

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

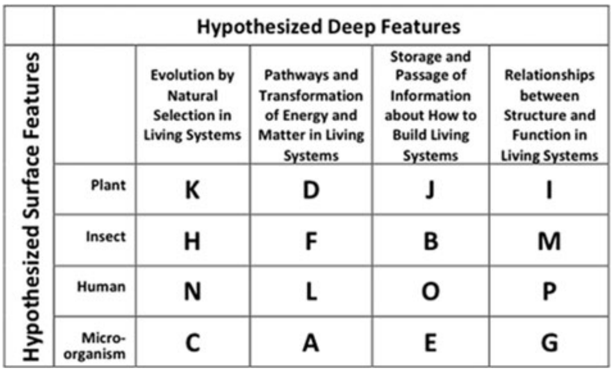
Cognitive and developmental psychologists have used a variety of research methods to measure and characterize expertise in chess, mathematics, and radiogram reading among radiologists, to name just a few examples (Chi, 2006b). A particularly promising methodological approach across all these studies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

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

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

METHODS

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



		Hypothesized Deep Features			
Hypothesized Surface Features		Evolution by Natural Selection in Living Systems	Pathways and Transformation of Energy and Matter in Living Systems	Storage and Passage of Information about How to Build Living Systems	Relationships between Structure and Function in Living Systems
	Plant	K	D	J	I
	Insect	H	F	B	M
	Human	N	L	O	P
	Micro-organism	C	A	E	G

Figure 1. Hypothesis-driven biology card stimulus set. Columns represent each of the four hypothesized deep features of biology, and rows represent each of the four hypothesized surface features of biology. Each letter represents one of sixteen biology problems that was printed on a card with that letter at the top. Each problem was crafted to contain both a single hypothesized surface feature and a single hypothesized deep feature.

Task Development

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

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

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

Task Conditions

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

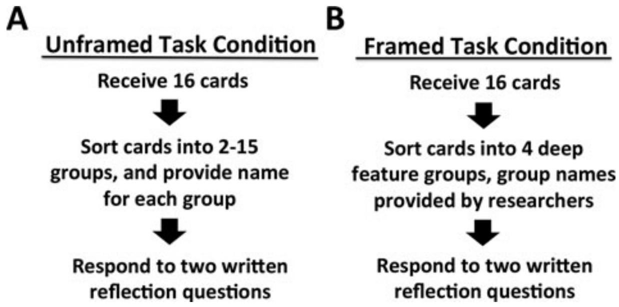

A
Unframed Task Condition
Receive 16 cards
Sort cards into 2-15 groups, and provide name for each group
Respond to two written reflection questions
B
Framed Task Condition
Receive 16 cards
Sort cards into 4 deep feature groups, group names provided by researchers
Respond to two written reflection questions

Figure 2. Unframed and framed task condition card sort protocols. (A) Unframed task condition protocol. (B) Framed task condition protocol. Participants always completed the unframed task condition before completing the framed task condition.

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

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

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

Analysis and Comparison of Constructed Card Groupings

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

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

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

Analysis and Comparison of Constructed Card Group Names

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

Analysis and Comparison of Responses to Reflective Prompts

A scoring rubric was also developed to assess the prevalence of sorting strategies based upon surface or deep features in a combined analysis of all the responses given by participants to the four reflection questions. In addition, we examined the prevalence of participants preferring the names that they gave to the card groupings (in the unframed condition) to those preferring the names provided by the researchers (in the framed condition).

Comparative Statistical Analyses

Two-tailed Student's *t* tests were used to compare the average percent card pairing and average edit distance measures between participant populations within a task condition. Comparisons of the results for a single participant population between the two task conditions—unframed and framed—were similarly analyzed. Additionally, we used a 2 (group: biology faculty, non–biology majors) by 2 (task: unframed, framed) repeated-measures analysis of variance (ANOVA) to examine the significance level of all of these comparisons. Results from *t* test analyses were confirmed by repeated-measures ANOVAs, so only the former are presented here. Pearson's chi-square tests were used to compare the prevalence of group names and specific card sorting strategies used by different participant populations. To normalize for differences in the size of particular participant populations, all variances are presented as an SE of the mean. All statistical comparisons were generated using JMP, version 9 (2010), or IBM SPSS, version 20.0 (2011).

Recruitment and Participant Population

Participants in this research were recruited from the students and faculty of a large urban university with more than 25,000 undergraduates (1800 biology majors and 5000 students enrolled in biology courses) and ~40 faculty members in biology who are active in research, as well as in teaching, and represent a wide breadth of subdisciplines spanning from the molecular to the ecological scale. Biology experts were recruited from among the tenured and tenure-track faculty. Biology novices were recruited from a general education course in biology populated primarily by non–biology major students during the first week of the semester. We hypothesized that non–biology majors enrolled in a biology course would have the greatest interest in and understanding of biology among the population of non–biology majors on campus, and thus we thought this population would be most representative of university-level biology novices. Each student in the course completed the tasks associated with this study as part of his or her course curriculum, but only those identified as non–biology majors were included in the study. Likewise, only those participants who completed tasks as directed by the researchers were included in the study. Each subject was allowed to decline participation without negative consequence. The committee for the protection of human subjects approved this research (protocol #X10-036).

RESULTS

This Biology Card Sorting Task yields multiple sources of data for analysis. To evaluate this task as a novel assessment tool, it is important to consider multiple measures of participants' performance. Below, a description of the participant populations is followed by example raw card sort data. Then, six analyses are presented that provide insights not only into how the two subject populations group the cards in the stimulus set, but also how these two populations named their constructed card groupings and rationalized their approach to the task. The six analyses presented are: 1) analysis of the number of card groups generated in the unframed task condition; 2) analysis of time to sort in both the unframed and framed task conditions; 3) calculation of prevalence of deep feature, surface feature, and unexpected card pairings in both task conditions; 4) calculation of edit distance from hypothesized deep feature and surface feature sorts in both task conditions; 5) analysis of prevalence of hypothesized deep feature and surface feature group names in the unframed task condition; and 6) analysis of card sorting rationales based on responses to the reflective questions. The figures, tables, and results are organized to show comparisons between non–biology majors and biology faculty and comparisons between the unframed and the framed card sort task conditions for each of these participant populations.

Description of Participant Populations

The participant populations for this biology card sort study are described in Table 1. From an invited pool of 35 tenured/tenure-track biology faculty, 23 participated for a 69% participation rate. From an invited pool of 131 undergraduate non–biology majors, all participated in doing the task as part of their class activity and the results for 101 are presented here for an actual 89% participation rate. Of the 30 undergraduates whose data are not included here, two were excluded because they were actually biology majors, 14 did not consent to have their data included in the study, and

Table 1. Participant population

Participant type	Number invited	Participation rate	Sample size	Female participants	Participants of color
Tenure-track biology faculty	35	69%	23	26%	48%
Undergraduate non–biology major	131	89%	101	55%*	55%

*$p = 0.014$ (Pearson chi-square).

14 had sorting anomalies (e.g., using a card twice or failure to use a card). There was a significantly greater proportion of females in the non–biology major population (55%) compared with the biology faculty population (26%; $\lambda^2 = 6.032$, $p = 0.014$). There was no significant difference in the proportion of participants of color between non–biology majors (55%) and biology faculty (48%). All analyses and comparisons described below are based on these non–biology majors ($n = 101$) and biology faculty ($n = 23$).

Example Card Sorts from a Non–Biology Major and a Biology Faculty for the Unframed and the Framed Task Conditions

Figure 3 shows example biology card sort task results from a single biology faculty member (Figure 3, A and B) and a single non–biology major student (Figure 3, C and D). These examples are shown to highlight two things. First, it is important to note that the biology card sorting task yields multiple sources of data for analysis, including the number of card groups generated, the membership and resulting card associations that result from these constructed groups, and the chosen name that an individual assigns to each group in the unframed task condition. Data from the reflective questions were also analyzed (raw data not shown). Second, these examples are shown to demonstrate that an exact hypothesized deep feature card sort (see columns in Figure 1) and an exact hypothesized surface feature card sort (see rows in Figure 1) were observed in the unframed task condition (see Figure 3, A and C, respectively). Of note, no biology faculty ever produced an exact hypothesized surface feature sort, nor did any non–biology major ever produce an exact hypothesized deep feature sort in the unframed task condition.

Analysis of Card Groupings and Resulting Surface Feature, Deep Feature, and Unexpected Card Pairings Constructed by Non–Biology Majors and Biology Faculty

As described in *Methods*, comparisons between constructed card groupings were accomplished by identifying all the card pairs that existed within a card group for each of the groups

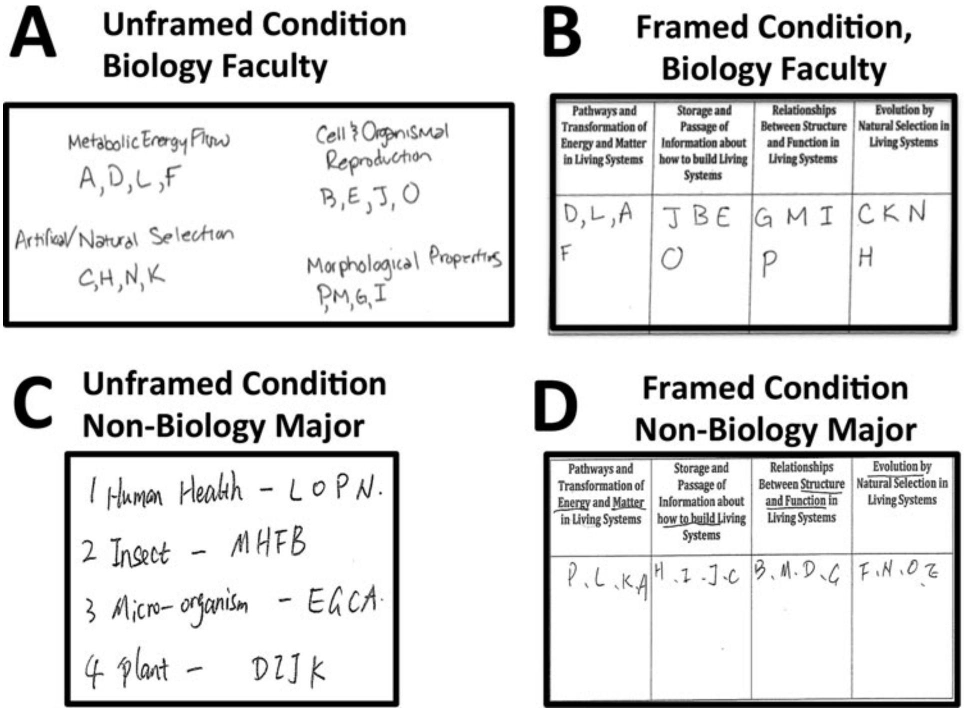

A
Unframed Condition
Biology Faculty
Metabolic Energy Flow
A, D, L, F
Cell & Organismal Reproduction
B, E, J, O
Artificial/Natural Selection
C, H, N, K
Morphological Properties
P, M, G, I
B
Framed Condition, Biology Faculty
Pathways and Transformation of Energy and Matter in Living Systems
Storage and Passage of Information about how to build Living Systems
Relationships Between Structure and Function in Living Systems
Evolution by Natural Selection in Living Systems
D, L, A F
J B E O
G M I P
C K N H
C
Unframed Condition
Non-Biology Major
1 Human Health - L O P N.
2 Insect - M H F B
3 Micro-organism - E G C A.
4 plant - D I J K
D
Framed Condition
Non-Biology Major
Pathways and Transformation of Energy and Matter in Living Systems
Storage and Passage of Information about how to build Living Systems
Relationships Between Structure and Function in Living Systems
Evolution by Natural Selection in Living Systems
P.L.K.A
H.I.J.C
B.M.D.G
F.N.O.E

Figure 3. Example card sort results. (A) Unframed task condition, biology faculty. (B) Framed task condition, biology faculty. (C) Unframed task condition, non–biology majors. (D) Framed task condition, non–biology majors. Note that the card groupings shown in (A) represent an exact hypothesized deep feature sort and the card groupings in (C) represent an exact hypothesized surface feature sort. Responses to reflection questions are not shown.

generated by an individual participant. Then, a percentage of hypothesized deep feature card pairs, hypothesized surface feature card pairs, and unexpected card pairs could be calculated for each individual's card sort. This approach was used to calculate average percentages of different types of card pairs for non–biology major and biology faculty populations, as well as to compare these proportions of different types of card pairs in the unframed and framed task conditions.

Percent Card Pairings in the Unframed Card Sort. In the unframed card sort condition (Figure 4A and Table 2), Biology faculty ($n = 23$) generated an average of 8.6 ± 2.2% surface feature card pairings, 71.7 ± 3.9% deep feature card pairings, and 19.8 ± 2.6% unexpected card pairings. In the unframed card sort, non–biology majors ($n = 101$) generated an average of 40.8 ± 2.9% surface feature card pairings, 29.2 ± 2.2% deep feature card pairings, and 30.0 ± 1.6% unexpected card pairings. Statistical comparison of these means showed that non–biology majors generated a significantly smaller average percentage of deep feature card pairings in the unframed task condition than biology faculty ($p < 0.0001$). In addition, non–biology major students generated a significantly greater average percentage of surface feature card pairings ($p = 0.0001$) and unexpected card pairings ($p = 0.0018$) in the unframed condition compared with biology faculty.

Percent Card Pairings in the Framed Card Sort. In the framed card sort condition (Figure 4B and Table 2), non–biology majors ($n = 101$) generated an average of 16.2 ± 0.8% surface feature card pairings, 39.6 ± 2.0% deep feature card pairings, and 44.2 ± 1.5% unexpected card pairings. In the framed card sort, biology faculty ($n = 23$) generated an average of 4.3 ± 0.8% surface feature card pairings, 83.1 ± 3.3% deep feature card pairings, and 12.6 ± 2.5% unexpected card pairings. Statistical comparison of these means showed that non–biology majors continued to generate a significantly smaller proportion of deep feature card pairings in the framed task condition than biology faculty ($p < 0.0001$). In addition, non–biology major students continued to generate a significantly higher proportion of surface feature card pairings ($p < 0.0001$) and unexpected card pairings ($p < 0.0001$) in the framed task condition compared with biology faculty.

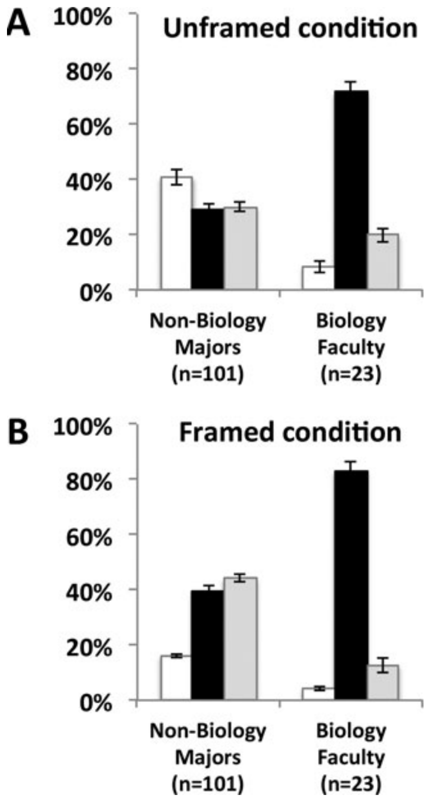


Figure 4. Average percentages of surface feature, deep feature, and unexpected card pairs. Average percentages of surface feature card pairs (white bars), deep feature card pairs (black bars), and unexpected card pairs (gray bars) generated in the unframed task condition (A) and the framed task condition (B) are shown for non–biology majors (left) and biology faculty (right). Comparison between non–biology majors and biology faculty showed that they construct significantly different proportions of each type of card pair compared with one another in both the unframed and framed task conditions ($p < 0.001$ or lower for each type of card pair). Additionally, non–biology majors constructed a significantly different proportion of each type of card pair in the framed task condition as compared with their results in the unframed condition ($p = 0.0006$ or lower for each type of card pair), whereas biology faculty only constructed a significantly different proportion of deep feature card pairs in the framed task condition as compared with their results in the unframed condition ($p < 0.03$). See *Results* for statistical details.

Comparison of Percent Card Pairings between the Unframed and Framed Card Sorts. Comparison of data from the unframed and framed conditions revealed significant shifts between the two task conditions within each participant population. In the framed task condition, non–biology majors constructed a significantly smaller proportion of surface feature pairs (16.2 ± 0.8% vs. 40.8 ± 2.9%; $p < 0.0001$) and a significantly greater proportion of deep feature (39.6 ± 2.0% vs. 29.2 ± 2.2%; $p = 0.0006$) and unexpected pairs (44.2 ± 1.5% vs. 30.0 ± 1.6%; $p < 0.0001$) compared with their initial results in the unframed task condition. Similarly, in the framed task, biology faculty constructed a significantly greater proportion of deep feature pairs (83.1 ± 3.3% vs. 71.7 ± 3.9%; $p < 0.0291$) and a somewhat smaller proportion of surface feature pairs

Table 2. Prevalence of surface feature, deep feature, and unexpected card pairings

		Unframed task condition			Framed task condition		
Participant type	n	Surface	Deep	Unexpected	Surface	Deep	Unexpected
Tenure-track 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)
Undergraduate non–biology major	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)

*$p < 0.001$ for comparisons between participant types on each measure shown above. SEM is in parentheses.

(4.3 ± 0.8% vs. 8.6 ± 2.2%; $p < 0.0849$) and unexpected pairs (12.6 ± 2.5% vs.19.8 ± 2.6%; $p < 0.0547$) compared with their initial results in the unframed task condition. In summary, both participant populations shifted toward a significantly greater proportion of deep feature card pairings, but only non–biology majors showed a significantly smaller proportion of surface feature card pairs and a significantly greater proportion of unexpected pairs in the framed task condition compared with the unframed task condition.

Analysis of Edit Distances from the ED-Surface and the ED-Deep Sorts

As described in *Methods*, comparisons between constructed card groupings by non–biology majors and biology faculty was also accomplished through a second analysis, in which an edit distance from the ED-Surface sort and an edit distance from the ED-Deep sort was calculated for each individual card sort. Edit distance is defined as the minimum number of card moves that would need to be made to turn an individual's card sort either into an exact hypothesized surface feature sort or an exact hypothesized deep feature. As such, an exact hypothesized deep feature sort would have an ED-Deep of 0 necessary card moves and ED-Surface of 12 necessary card moves. Similarly, an exact hypothesized surface feature sort would have an ED-Deep of 12 necessary card moves and ED-Surface of 0 necessary card moves. From these calculated edit distances, an average ED-Surface and ED-Deep could be calculated and compared for the non–biology major population and the biology faculty population, as well as between the unframed and framed task conditions for each population. Units of ED are "necessary card moves."

Edit Distance in the Unframed Card Sort. In the unframed card sort condition (Figure 5A and Table 3), biology faculty ($n = 23$) constructed card sorts with an average ED-Surface of 10.9 ± 0.3 and an average ED-Deep of 4.5 ± 0.5. Non–biology majors ($n = 101$) constructed card sorts with an average ED-Surface of 7.0 ± 0.3 and an average ED-Deep of 8.2 ± 0.3. Statistical comparison of these means showed that non–biology majors and biology faculty were significantly different from one another ($p < 0.0001$).

Edit Distance in the Framed Card Sort. In the framed card sort condition (Figure 5B and Table 3), non–biology majors ($n = 101$) constructed card sorts with an average ED-Surface of 9.1 ± 0.1 and an average ED-Deep of 5.9 ± 0.3. Biology faculty ($n = 23$) constructed card sorts with an average ED-Surface of 11.0 ± 0.2 and an average ED-Deep of 1.2 ± 0.3. Statistical comparison of these means showed that non–biology majors and biology faculty were significantly different from one another ($p < 0.0001$).

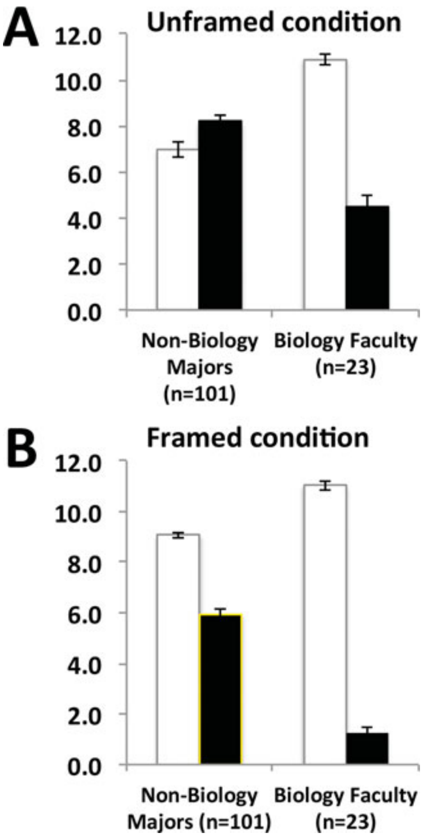


Figure 5. Edit distances from ED-Surface and ED-Deep. Calculated ED-Surface (white bars) and ED-Deep (black bars) for the unframed task condition (A) and the framed task condition (B) are shown for non–biology majors (left) and biology faculty (right). Comparison between non–biology majors and biology faculty showed significantly different average ED-Surface and average ED-Deep compared with one another in both the unframed and framed task conditions ($p < 0.0001$ for each comparison). Additionally, non–biology majors had significantly different average ED-Surface and average ED-Deep in the framed task condition as compared with their results in the unframed condition ($p = 0.0001$), whereas biology faculty only had significantly ED-Deep in the framed task condition as compared with their results in the unframed condition ($p < 0.0001$). See *Results* for statistical details.

Comparison of Edit Distance between the Unframed and Framed Card Sorts. Comparison of ED-Surface and ED-Deep analyses for the unframed and framed task conditions

Table 3. Edit distances (ED) from the hypothesized surface feature sort and the hypothesized deep feature sort[a]

		Unframed task condition		Framed task condition	
Participant type	n	ED from surface sort	ED from deep sort	ED from surface sort	ED from deep sort
Tenure-track biology faculty	23	10.9 (0.3)	4.5 (0.5)	11.0 (0.2)	1.2 (0.3)
Undergraduate non–biology major	101	7.0* (0.3)	8.2* (0.3)	9.1* (0.1)	5.9* (0.3)

*$p < 0.0001$ for comparisons between participant types on each measure shown above. SEM is in parentheses.

[a]Note that lower ED numbers indicate sorts more similar to the hypothesized sort.

Table 4. Rubric to quantify prevalence of hypothesized deep feature group names in the unframed sort

Category title of the hypothesized deep features of the card sort	Titles generated by subjects in the unframed sort that were accepted as equivalent to those given by researchers in the framed sort
Evolution by natural selection in living systems	Any title using the term *evolution* Genetic selection Natural selection Survival of the fittest Adaptation Speciation Evo-devo
Pathways and transformations of energy and matter in living systems	Any title using the term *energy* Nutrient production and use Metabolism Cellular respiration
Storage and passage of information about how to build living systems	Inheritance Heritability Genetics DNA/traits DNA: the genetic blueprint for organisms
Relationships between structure and function in living systems	Form and function Parts and function Structure relates to function Refers to the function of a structure or part Functional morphology Biomechanics

revealed significant shifts between the two task conditions for each participant population. In the framed task, non–biology majors constructed card sorts that had ED-Surface (9.1 ± 0.1) that were statistically further away from an exact hypothesized surface feature sort compared with their constructed sorts in the unframed task condition (7.0 ± 0.3; $p < 0.0001$). In addition, non–biology majors constructed card sorts in the framed task condition that had ED-Deep (5.9 ± 0.3) that were significantly closer to an exact hypothesized deep feature sort compared with their initial results in the unframed task condition (8.2 ± 0.3; $p < 0.0001$). In the framed task, biology faculty constructed card sorts that had ED-Surface (11.0 ± 0.2) that were statistically indistinguishable from their constructed sorts in the unframed task condition (10.9 ± 0.3; $p = 0.6827$). However, biology faculty constructed card sorts in the framed task condition that had ED-Deep (1.2 ± 0.3) that were significantly closer to an exact hypothesized deep feature sort compared with their initial results in the unframed task condition (4.5 ± 0.5; $p < 0.0001$). In summary, both participant populations shifted toward constructing card groupings that were more similar to the exact hypothesized deep feature sort in the framed task condition compared with the unframed task condition, with non–biology majors also significantly shifting away from constructing card groupings closer to the exact hypothesized surface feature sort.

Analysis and Comparison of Constructed Card Group Names

While the analyses presented above provide insights into how the participant populations grouped the cards in the stimulus set, the analyses below describe how these participant populations chose to name their constructed card groupings in the unframed task condition. As described in *Methods*, comparisons between the names given to constructed card groupings by non–biology majors and biology faculty were accomplished through blind coding of the card group names for the presence of hypothesized deep features or hypothesized surface features. Two observers analyzed all card group names. Groups names that included the terms *human*, *plant*, *insect*, or *microorganism* were coded as representing each of those hypothesized surface features, respectively. For deep features, it was necessary to develop a rubric more specifically defining the terminology found in group names that would be coded as representing each hypothesized deep feature (see Table 4). These rubrics were developed and refined based on responses seen in the data set, examples of which are listed in the second column. Interrater reliability for each analysis presented was greater than 95% agreement between observers.

Hypothesized Deep Feature Group Names. Analysis of the prevalence of card group names related to the four hypothesized deep features (see columns in Figure 1) is shown in Figure 6. For all four hypothesized deep features, a significantly larger proportion of biology faculty ($n = 23$) used each hypothesized deep feature in naming one or more of their card groups as compared with non–biology majors ($n = 101$). More specifically, the deep feature "Evolution by Natural Selection in Living Systems" appeared in the group names of a significantly larger proportion of biology faculty (87.0%) as compared with non–biology majors (19.8%; $\lambda^2 = 38.7$, df = 1, $p < 0.0001$; Figure 6A). The deep feature "Pathways and Transformations of Energy and Matter" also appeared in the group names of a significantly higher proportion of biology faculty (82.6%) as compared with non–biology majors (8.9%; $\lambda^2 = 58.2$, df = 1, $p < 0.0001$; Figure 6B). The deep feature "Storage and Passage of Information about How to Build Living Systems" appeared in the group names of a significantly larger proportion of biology faculty (95.7%) as compared with non–biology majors (39.6%; $\lambda^2 = 23.5$, df = 1, $p < 0.0001$; Figure 6C). Finally, the deep feature "Relationships between Structure and Function in Living Systems" appeared in the group names of 39.1% of biology faculty as compared

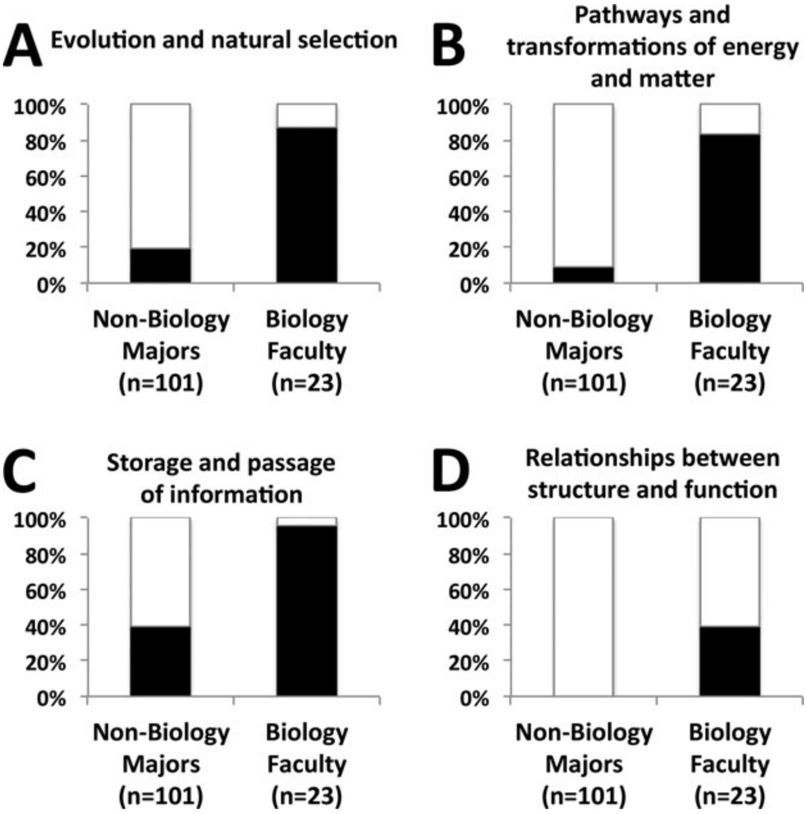


Figure 6. Prevalence of deep features in card group names in the unframed task condition. Proportions of participants who included (black bars) or did not include (white bars) each deep feature in one or more of their card group names in the unframed task condition are shown for non–biology majors (left) and biology faculty (right) for each of the hypothesized deep features: (A) evolution and natural selection (B) pathways and transformations of energy and matter, (C) storage and passage of information, and (D) relationships between structure and function. A significantly larger proportion of biology faculty included each deep feature in their card group names as compared with non–biology majors ($p < 0.0001$ for each deep feature). See *Results* and Table 4 for analysis rubrics, and *Results* for statistical details.

with none of the non–biology majors (0%; $\lambda^2 = 42.6$, df = 1, $p < 0.0001$; Figure 6D).

Hypothesized Surface Feature Group Names. Analysis of the prevalence of card group names related to the four hypothesized surface features (see rows in Figure 1) is shown in Figure 7. For all four surface features, a significantly larger proportion of non–biology majors ($n = 101$) used each hypothesized surface feature in naming one or more of their card groups as compared with biology faculty ($n = 23$). More specifically, the surface feature "Human" appeared in the group names of a significantly larger proportion of non–biology majors (47.5%) as compared with biology faculty (8.7%; $\lambda^2 = 11.7$, df = 1, $p = 0.0006$; Figure 7A). The surface feature "Insect" appeared in the group names of 43.6% of non–biology majors as compared with none of the biology faculty (0%; $\lambda^2 = 15.5$, df = 1, $p < 0.0001$; Figure 7B). The surface feature "Plant" appeared in the group names of a significantly larger proportion of non–biology majors (51.5%) as compared with biology faculty (17%; $\lambda^2 = 8.8$, df = 1, $p = 0.003$; Figure 7C). Finally, the surface feature "Microorganism" appeared in the group names of 36.6% of non–biology majors as compared with none of the biology faculty (0%; $\lambda^2 = 12.0$, df = 1, $p = 0.0005$; Figure 7D).

Analysis and Comparison of Card Sorting Strategy Explanations from Responses to Posttask Reflection Questions

In addition to analyzing how participants grouped the cards and named these groups, we also analyzed participants' reported card sorting strategy explanations, which appeared in their responses to posttask reflection questions. As described in *Methods*, analysis of participants' card sorting strategies was accomplished through blind coding of participants' reflection narratives. In particular, narratives were coded for rationales that included explicit reference to using either hypothesized surface features or hypothesized deep features in sorting. Two observers analyzed all narrative responses to the reflection questions, and the interrater reliability for each analysis presented was greater than 94% agreement between observers.

Analysis of the participants' card sorting strategy explanations is shown in Table 5. Sample quotes from both biology faculty and non–biology majors who evidenced each rationale are shown. In their reflection narratives, 100% of biology faculty made reference to one or more of the four hypothesized deep features as part of their sorting strategy as compared with only 22% of non–biology majors ($\lambda^2 = 47.9$, df = 1, $p < 0.0001$). In contrast, only two faculty (8.7%) made reference to using hypothesized surface features in their sorting, as compared with 37.6% of non–biology majors ($\lambda^2 = 7.2$, df = 1, $p = 0.0074$).

Analysis of Number of Card Groups Generated in Unframed Task Condition

The average number of card groups generated in the unframed task by non–biology majors (5.3 ± 0.2, $n = 101$) was significantly fewer than the number generated by biology faculty (6.5 ± 0.3, $n = 23$; $p = 0.0011$).

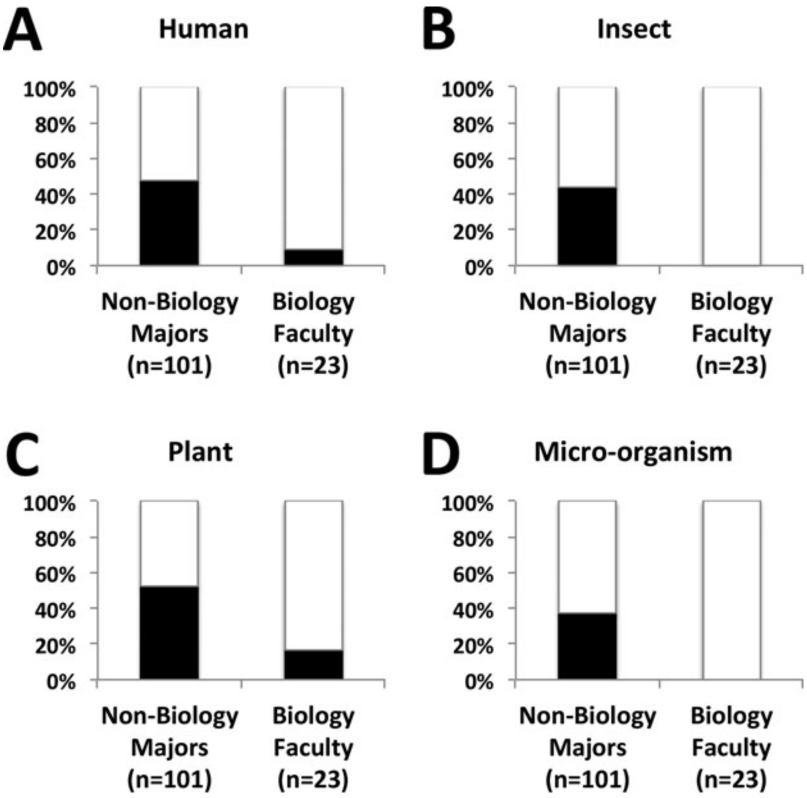


Figure 7. Prevalence of surface features in card group names in the unframed task condition. Proportions of participants who included (black bars) or did not include (white bars) each surface feature in one or more of their card group names in the unframed task condition are shown for non–biology majors (left) and biology faculty (right) for each of the hypothesized surface features: (A) "Human," (B) "Insect," (C) "Plant," and (D) "Microorganism." A significantly larger proportion of non–biology majors included each surface feature in their card group names as compared with biology faculty ($p < 0.003$ or lower for each surface feature). See *Results* for analysis rubric and statistical details.

Average Completion Times for the Unframed and Framed Task Conditions

The average time taken to complete the unframed task condition by non–biology majors (12.0 ± 0.5 min, $n = 96$) was significantly shorter ($p = 0.0547$) than for biology faculty (15.1 ± 1.5 min, $n = 23$). However, the average time taken to complete the framed task condition by non–biology majors (6.8 ± 0.3 min, $n = 93$) was significantly longer ($p = 0.0001$) than for biology faculty (4.7 ± 0.4 min, $n = 23$). Comparison of task completion time data within each participant population for the unframed and framed conditions revealed significantly faster task completion times for both non–biology majors (12.0 ± 0.5 vs. 6.8 ± 0.3 min; $p < 0.0001$) and biology faculty (15.1 ± 1.5 vs. 4.7 ± 0.4 min; $p < 0.0001$) in the framed task condition.

Table 5. Rubric for and analysis of card sorting strategy explanations

Participant type	Surface feature rationale	Sample quote
Tenure-track biology faculty ($n = 23$)	8.7%	"Others [cards] are united by the kind of organism (DIPMG)." "Plant, well plant are plants and they are just odd."
Undergraduate non–biology major ($n = 101$)	37.6%	"I grouped certain cards together because I looked for key words in the problems such as: insects, humans, cells, and plants." "With plants, any reference to photosynthesis, reproduction, environment, would categorize them into such a group."
	Deep feature rationale	
Tenure-track biology faculty ($n = 23$)	100%	"1st group dealt w/ manner in which all organisms (microbe, plant or animal) process energy within their cells, energy metabolism was a unifying theme." "Evolution and natural selection make sense to me but in another level I also liked structure and function. As for L, on the surface it seems not to fit but when you attempt to answer the question, energy and matter seemed the best category.
Undergraduate non–biology major ($n = 101$)	22%	"Bacteria & Pesticides evolving to battle their 'cures' was categorized as evolution." "Energy seemed to be an underlying theme (atp) for a few."

DISCUSSION

While there are widespread aspirations to focus undergraduate biology education on teaching students to think like biologists, which includes development of students' conceptual expertise in biology, there is a dearth of assessment tools designed to measure progress from more novice biological thinking toward more expert biological thinking. In this paper, we present a novel assessment tool, the Biology Card Sorting Task, which has been adapted from card sorting approaches used in cognitive psychology research (Chi *et al.*, 1981). This task has been designed to probe how an individual organizes his or her biological conceptual knowledge with the aim of describing and distinguishing the organization of biological knowledge in putative biological experts (biology faculty) and putative biological novices (non–biology majors). In the following sections, we describe insights gained about the task from comparing experimental results from biology faculty and non–biology majors, how this assessment tool differs from other biology conceptual assessment tools, how this card sorting task is unique compared with previously developed card sorting tasks, and we explore future directions for the adaptation and use of this novel assessment task more broadly.

The Biology Card Sorting Task Distinguishes Putative Experts and Novices—Differences in How Non–Biology Majors and Biology Faculty Perform on the Task

Biology faculty and non–biology majors differed significantly on each of the six analytical measures for the Biology Card Sorting Task used to compare the card sorts produced by these two groups. These data confirm that this novel assessment task appears to be robust in distinguishing populations of putative biology experts and putative biology novices on a variety of measures. In addition, this initial investigation also revealed multiple interesting findings about the differences in card sorting results for these two populations. Similar to previous studies, biology faculty took more time to complete their sorts than did non–biology majors (Chi *et al.*, 1981). In addition, biology faculty constructed on average one more card group than non–biology majors. More specific results for each population are considered below.

Biology Faculty Appear to Sort Based on Deep Features. Strikingly, biology faculty grouped cards together in a manner we hypothesized experts would, grouping cards primarily based on deep features (fundamental biological principles). This conclusion is supported by the high proportion of deep feature card pairings produced in the unframed task condition by biology faculty (Figure 4A and Table 2). In addition, the observation that biology faculty had average card sort edit distances closer to the hypothesized deep feature sort than the hypothesized surface feature sort is additional supporting evidence (Figure 5A and Table 3). That biology faculty members are sorting based on deep features is also supported by their sorting strategies, with 100% of faculty naming one or more of the deep features as a part of their sorting strategy (Table 5), and by choices in naming the card groups that they create, explicitly choosing group names that aligned well with the four hypothesized deep feature categories (see *Results* and Figure 6). All these data support the hypothesis that biology faculty members appear to be primarily using deep features—namely core biological concepts—to sort the biology problems on the cards.

Non–Biology Majors Do Not Necessarily Sort Based on Surface Features. As hypothesized, non–biology majors did not appear to group cards based primarily on deep features; however, they also did not appear to group cards based on surface features, as we had hypothesized they would. This was evidenced by the presence of relatively comparable proportions of surface feature, deep feature, and unexpected card pairings (Figure 4A and Table 2), as well as by the fact that non–biology majors produced card sorts in the unframed condition with edit distances that were almost equidistant from both the hypothesized novice and expert card sorts (Figure 5A and Table 3). In examining how non–biology majors named their card groups, 40% of non–biology majors used language aligned with the four surface features (see *Results* and Figure 7) to describe their card groupings, whereas a range of 0–40% used the four deep feature group names (see *Results* and Figure 6). These data suggest that the non–biology major population may be sorting using a variety of different organizational frameworks, not just an organismal framework. Multiple hypotheses could explain the variation seen among non–biology majors in this task. First, non–biology majors are a less homogeneous population by nature than faculty. Some non–biology majors may have intermediate biology knowledge and may be performing this assessment task like partial experts, whereas others may have had minimal exposure to biology and may even struggle with the language of the cards. Even though we attempted to minimize biological jargon on the card stimulus set, there were comments by a subset of non–biology majors that some language on a few of the cards was challenging (e.g., bilateral symmetry; stratified epithelium [unpublished data]). In future versions of the card stimulus set, we would further translate or remove that biological language identified by participants in this study as challenging. Finally, published card sorting results among novices across a variety of fields of expertise tend to exhibit wider variation than results found among experts (Chi *et al.*, 1981; Chi, 2006a; Mason and Singh 2011). A future study to conduct think-aloud interviews with non–biology majors may provide additional insights into the variety of organizational frameworks this particular population is using to perform the assessment task.

Biological Framing Sharpens Biology Faculty Sorts toward Deep Features and Shifts Non–Biology Major Sorts Away from Surface Features. When we compare the card sorting results between the unframed and framed task conditions, biology faculty appear to sort very similarly, with some sharpening of their use of a hypothesized deep feature framework. Biology faculty produced more deep feature card pairings and fewer surface feature and unexpected card pairings in the framed task condition (Figure 4 and Table 2); however, only the increase in deep feature pairs was statistically significant. Similarly, the biology faculty edit distance from the hypothesized expert sort was reduced to only ~1 card move difference (Figure 5 and Table 3). This suggests that the explicit biological framing and requirement to sort into the researcher's four deep feature categories in the framed task condition appeared to slightly sharpen but not fundamentally alter the biology faculty population's performance on this task.

In contrast, the framed task condition did not appear to primarily shift non–biology majors' results toward a more expert framework based on deep features. As a population, non–biology majors did, however, exhibit a dramatic reduction by 24.6% in the proportion of surface feature card pairings they constructed, with a larger increase in the proportion of unexpected card pairings (14.2% increase) compared with the increase in deep feature card pairings (10.4% increase; Figure 4 and Table 2). This suggests that non–biology majors may have recognized that a surface feature approach was not possible in the framed task condition, but they were unable to make use of the given deep feature group names. Similarly, examination of the edit distance metric shows that non–biology majors shifted away from the hypothesized surface feature sort, and by definition, then, toward the hypothesized deep feature sort. However, this shift was from an ~8 card move difference from a deep sort feature sort in the unframed task condition to an ~6 card move difference from a hypothesized deep feature sort in the framed task condition (Figure 5 and Table 3). Taken together with the data on shifts in proportion of card pairings, these data suggest that non–biology major performance in the framed task condition may reflect their abandonment of the surface feature organizational framework as opposed to an active shift toward more deep feature and expert-like sorting.

Both Non–Biology Majors and Biology Faculty Generate Unexpected Card Pairings. Both participant populations produced unexpected card pairings, namely card pairings that neither represented a hypothesized surface feature pair (e.g., NL, two human cards; Figure 1) nor a hypothesized deep feature pair (e.g., KH, two evolution cards; Figure 1). The proportion of unexpected card pairings was always lower for biology faculty than non–biology majors (Figure 4 and Table 2). However, it is intriguing that the proportion of unexpected card pairings increased between the unframed and the framed task condition for non–biology majors, whereas this proportion decreased for biology faculty (Figure 4 and Table 2). One possible explanation of this result is that, as described above, the framed task condition and the explicitly given deep feature categories were mostly inaccessible and not understood by the non–biology majors, in contrast to biology faculty. This would lead to the conclusion that the non–biology majors population as a whole did not organize their biological ideas with respect to fundamental biological ideas and that they were unable to recognize and use the organizational framework of fundamental biological ideas given to them in the framed task condition. This also suggests that this particular population of undergraduate students did not have a partial expert framework that could be revealed by explicitly showing them deep feature category names in the framed task condition.

Another key observation was that biology faculty and non–biology majors produced different types of unexpected card pairings. Biology faculty commonly produced unexpected card pairings that were the result of pairing of cards associated with the two particular deep features: "Evolution by Natural Selection in Living Systems" and "Storage and Passage of Information about How to Build Living Systems" (unpublished data). While pairing of cards from these two different deep feature categories would be characterized as an unexpected pair in our analyses, these ideas are intimately connected in biology, with the inheritance of genetic information being the substrate upon which evolutionary mechanisms act.

In contrast, the unexpected pairings produced by non–biology majors were less predictable. No unexpected card pairings were particularly more prevalent than any other. Sometimes these unexpected card pairings appeared to reflect a known misconception. One such example that was seen was the grouping of a card about fungi together with several plant cards (unpublished data). Further detailed analysis of the unexpected card pairings produced by non–biology majors may reveal more patterns.

Unique Aspects of the Biology Card Sorting Task Compared with Other Biology Conceptual Assessment Tools

The Biology Card Sorting Task presented here is intended to expand the repertoire of biological assessment tools available both to researchers from multiple disciplines and to practitioners ranging from individual instructors to departments and larger collaborative initiatives. The Biology Card Sorting Task is unique compared with other currently available biology assessment tools in at least two key ways.

Probing Connections between Conceptual Ideas Rather Than the Ideas Themselves. The Biology Card Sorting Task was designed to assess conceptual expertise in biology—namely how individuals organize their knowledge of biology and how they connect or do not connect biological ideas. While a variety of tools to assess biological conceptual knowledge have been developed, there does not appear to be a tool currently available that probes how individuals organize their biological knowledge broadly across multiple biological ideas. Many biology assessment tools have been developed that probe an individual's particular knowledge of a specific biological concept such as evolution (e.g., Anderson *et al.*, 2002; Baum *et al.*, 2005, Nehm and Schonfeld, 2008), molecular biology and genetics (e.g., Bowling *et al.*, 2008; Smith *et al.*, 2008; Shi *et al.*, 2010), host–pathogen interactions (e.g., Marbach-Ad *et al.*, 2009), osmosis and diffusion (e.g., Odom and Barrow, 1995; Fisher *et al.*, 2011), and energy and matter in living systems (e.g., Wilson *et al.*, 2007; Hartley *et al.*, 2011), to name just a few. Assessment tools such as these have been designed primarily to measure the presence or absence of particular pieces of knowledge or particular misconceptions about that knowledge, rather than the overall structure and organization of an individual's biological knowledge. More recently, some researchers have begun to employ these and other tools to explore connections students are making between ideas in a particular concept in biology (e.g., Wilson *et al.*, 2007; Nehm and Schonfeld, 2008; Hartley *et al.*, 2011). The Biology Card Sorting Task may, however, be one of the first tools developed specifically to assess an individual's conceptual expertise in biology, namely the organization of his or her biological understanding, across a wide range of conceptual ideas in biology.

Assessing How Individuals Will Perform on a Task versus Selection of an Answer Choice. The Biology Card Sorting Task presented here is unique in how an individual's thinking about biology is probed. Few biology concept assessment tools currently in use are open-ended, requiring individuals to produce evidence of their thinking through either writing or speaking (e.g., Wilson *et al.*, 2007). The majority of currently available assessment tools are closed-ended in their format, and students are asked to select their answer preference from a list of four to five given answers (e.g., Klymkowsky *et al.*, 2003; Semsar *et al.*, 2011). As such, while these assessment tools provide insight into some aspects of student thinking, many other aspects of student thinking are likely going undetected. Dissociations between students' success in selecting a scientifically accurate answer and an independent analysis of their ability to provide a scientific explanation for that answer choice have been previously documented (Rabinowitz and Mohammadreza, 1989; Anbar, 1991; Bridgeman and Morgan, 1996). As a complex task that involves sorting and writing, the Biology Card Sorting Task is an open-ended assessment tool that demands a different type of behavioral output from students compared with other available assessment tools, perhaps providing insight into new aspects of student thinking and conceptual organization that are not currently being monitored or explored.

Intentionally Unique Aspects of the Biology Card Sorting Task Compared with Other Cognitive Psychology Card Sorting Tasks

While the novel Biology Card Sorting Task presented here was inspired by published card sorting tasks from the developmental and cognitive psychology research literatures, it was intentionally designed to differ from and improve upon previously developed card sorting tasks in four key ways. These design differences were driven by previously described limitations of card sorting tasks generally, as well as by the desire to develop a task-based assessment tool that would be adaptable and feasible for broad use.

Hypothesis-Driven Card Stimulus Set. First, the Biology Card Sorting Task was conceptually hypothesis-driven in its design. Previously described card sorting tasks have either not explicitly described why particular card stimuli were selected for use or implied that card stimuli were somewhat randomly selected. In contrast, this Biology Card Sorting Task and its card stimulus set were specifically designed to test a particular hypothesis about how novices might organize their ideas in biology—in terms of organism type (see row identifiers in Figure 1)—and a specific hypothesis about how experts might organize their ideas in biology—in terms of fundamental biological concepts (see column titles of Figure 1). The purposeful development of a hypothesis-driven card set was nontrivial, requiring each card to be able to be seen in the context of both of these two hypothesized conceptual frameworks, as well as to be in language that was as accessible as possible to both novices and experts. Importantly, this approach to constructing a card stimulus set for the Biology Card Sorting Task is now ripe for adaptation and the development of new card stimulus sets that could test different hypotheses about how individuals organize their biological knowledge.

Probing Individuals' Own Conceptual Frameworks, as Well as Their Interpretation of Expert Frameworks: Using Two Card Sort Task Conditions with and without Biological Framing. The Biology Card Sorting Task was also designed to allow individuals both to share their own initial approach to organizing biological ideas (unframed task condition) and to attempt to navigate one possible biological expert framework (framed task condition). Previous card sorting tasks were primarily open-ended, with participants sorting cards into as many groups as they liked (Chi, 2006a,b). In some cases, participants were allowed to do multiple, open-ended sorts with the same card set, referred to as serial repetitive card sorting. However, these tasks did not explicitly probe whether individuals might be able to use an expert framework if explicitly cued to do so.

Because the development of expertise in any discipline is complex, it is no doubt a process that may involve stages. As such, these two (or eventually more) task conditions may be particularly important when studying emerging experts, such as biology majors, and identifying a point at which they may be unable to construct a hypothesized expert framework but may be able to recognize and apply an expert framework. For example, individuals with little to no expertise in biology may be able to come up with some organizational framework for the cards in the unframed task, but then are not be able to make sense of an expert framework presented to them in the framed task. In contrast, there may also be individuals with incomplete or intermediate conceptual expertise, such as undergraduate biology majors who may not initially apply an expert framework on their own in the unframed task, but who would be able to navigate and apply an expert framework if explicitly presented with it. To our knowledge, this Biology Card Sorting Task is unique in having the potential to be able to discern putative stages of the development of biological conceptual expertise by assessing both students' initial sorting strategies and then their actions in response to being given an expert biological framework.

Multiple Approaches to Quantitative and Qualitative Analysis of Card Sorts. The data sets generated by conducting the Biology Card Sorting Task are large and complex, including the card groups themselves, the names for the card groups, and the individuals' narrative responses about their sorting strategies for the two task conditions. Previously published card sorting studies have generally reported results using a nonsystematic and nonquantitative case study method in which an individual's approach is described and presented (Chi *et al.*, 1981). In a few cases, investigators have quantified some aspects of card sorting results, such as the percent of individuals who assigned a particular card to a particular category, but still with small numbers of participants (reviewed in Chi, 2006a; Mason and Singh, 2011). Given our desire to make this assessment tool useful with large groups of students, we have also developed companion analytical techniques that enable automated, quantitative analysis of card groupings, as well as rubrics to enable blind scoring of the qualitative data that are generated as card group names and rationales for sorting strategies. As such, data entry of card groups generated by an individual is automatically analyzed via a Python computer script to generate quantitative metrics—percent hypothesized surface feature card pairs, percent deep feature card pairs, percent unexpected card pairs, edit distance from a hypothesized novice sort, and edit distance from a hypothesized expert sort—for each individual's produced card sorts in both the unframed and the framed task conditions. Analysis of the qualitative data generated from card group names and rationales for sorting strategies is not currently automated; however, rubrics exist for coding of these data. The Biology Card Sorting Task may afford investigators the opportunity to be both systematic and quantitative in analyzing the rich data set that emerges from conducting these assessment tasks with large numbers of individuals.

A Tool for Classroom Assessment, Program Assessment, and Biology Education Research. The Biology Card Sorting Task presented here offers the benefit of using a complex, open-ended task to assess the development of biology conceptual expertise that could be used in both research laboratories and classrooms. One aspiration for the Biology Card Sorting Task was to develop a novel task, grounded in the theoretical frameworks of cognitive psychology, that could also be used by biology departments with large numbers of students to provide insight into the progression of student conceptual thinking in biology over the course of an undergraduate degree program. Previous card sorting tasks and protocols that have been developed to study expertise have largely been designed for individual, think-aloud interview protocols, which by nature limit their use to research laboratories. In addition, it is often argued that concept inventories are chosen as an assessment tool due to the ease of quickly analyzing data that emerges from such a closed-ended tool. While many acknowledge that open-ended assessment tools capture a richer profile of student thinking, the resulting data are simply too complex and time-consuming to analyze. The Biology Card Sorting Task is an attempt to develop an approach to measuring the development of discipline-based conceptual expertise that generates a rich profile of student thinking on a complex task, while also generating data that are feasible to analyze at the level of classroom assessment, program assessment, and in the context of biology education research.

Characteristics of This Card Stimulus Set and Potential Adaptations

The Biology Card Sorting Task presented here is a first attempt to translate a promising measurement methodology from the cognitive sciences into an assessment tool that can be used in biology education and research. The card stimulus set presented here, however, is far from the only possible card stimulus set that could be used in measuring the development of biological expertise. In addition, the general structure of this card sorting, task-based assessment could be useful in measuring a variety of aspects of biological expertise in a variety of subdomains of this discipline, as well as across science disciplines.

The card stimulus set presented here was designed to test the specific hypothesis that putative biological experts would connect and organize biological ideas using an organizational framework consisting of four specific deep features: evolution, structure–function, information flow, and transformations of energy and matter. While these four core ideas align rather well with recently published frameworks attempting to delineate core ideas in the biological sciences (AAAS, 2011; College Board, 2013), the alignment is by no means exact. For example, the *Vision and Change* document (AAAS, 2011, p. 12), proposes five core concepts for biological literacy that consist of the four deep features we have used here and an additional core concept entitled *systems.* Similarly, the revised organizational framework for AP Biology recently put forward by the College Board includes four core concepts: evolution, genetics and information transfer, cellular processes: energy and communication, and interactions (College Board, 2013). Two of these core ideas align well with two deep features of the Biology Card Sorting Task (e.g., evolution, and genetics and information transfer); the third core idea—cellular processes: energy and communication—may be somewhat more expansive than our transformations of energy and matter deep feature. The core idea of interactions is similar to the systems core idea in *Vision and Change* and could be a separate deep feature in a future card stimulus set. These comparisons are relevant, because the currently used card stimulus set may need further refinement to be maximally useful in some educational settings. Our own results suggest that the addition of a systems or interaction deep feature category to the current card stimulus set would bring the Biology Card Sorting Task into better alignment with some policy documents (AAAS, 2011).

Additionally, this initial card stimulus set was designed to test the hypothesis that putative biological novices would connect and organize biological ideas using the framework of four specific surface features based on organism type: "Human," "Plant," "Insect," and "Microorganism." While this hypothesized novice framework for organizing biological knowledge appeared to capture ~40% (Figure 7 and Table 5) of the non–biology majors approaches to sorting, a majority of the non–biology majors appear to be sorting using organizational frameworks that we are not yet able to characterize. Adaptation of the current card stimulus set to test other hypothetical frameworks that novices may be using to organize their biological knowledge would produce another variation of the Biology Card Sorting Task.

Finally, multiple agencies are calling for more interdisciplinary courses and programs in undergraduate science education, yet there are few assessment tools available to investigate the extent to which students are connecting knowledge across the disciplines (AAMC-HHMI, 2009). An adaptation of this Biology Card Sorting Task could produce a task-based assessment tool to probe whether students in interdisciplinary courses and programs are more likely than students in traditional courses to sort problems into surface features, such as chemistry, physics, and biology, versus deep features, such as energy and matter and structure and function. The novel structure and quantitative analytical approaches of the Biology Card Sorting Task may be useful to discipline-based science education researchers across the sciences to test an infinite number of hypotheses about how a variety of different populations organize their conceptual knowledge within and across the scientific disciplines.

CONCLUSIONS

In conclusion, we have developed a novel assessment tool for the biological sciences that moves away from assessing individual pieces of knowledge and moves toward measuring how individuals organize biological ideas and develop biological conceptual expertise. This initial investigation of the Biology Card Sorting Task demonstrated that biology faculty and non–biology majors differed significantly on each of six analytical measures used to compare the card sorts produced by these two groups. As such, these data confirm that this novel assessment task appears to be robust in distinguishing populations of putative biology experts and putative biology novices on a variety of measures. With this shown, investigations of how individuals with various levels of biology experiences perform on the Biology Card Sorting Task may now be conducted. Finally, the general structure of the Biology Card Sorting Task may be adaptable for use in assessing other aspects of developing expertise in the biological sciences and beyond.

REFERENCES

American Association for the Advancement of Science (2011). Vision and Change in Undergraduate Education: A Call to Action, Washington, DC. http://visionandchange.org/files/2011/03/Revised-Vision-and-Change-Final-Report.pdf (accessed 22 May 2013).

Anbar M (1991). Comparing assessments of students' knowledge by computerized open-ended and multiple-choice tests. Acad Med *66*, 420.

Anderson DL, Fisher KM, Norman JG (2002). Development and validation of the conceptual inventory of natural selection. J Res Sci Teach *39*, 952–978.

Association of American Medical Colleges and Howard Hughes Medical Institute (2009). Scientific Foundations for Future Physicians: Report of the AAMC-HHMI Committee, Washington, DC: AAMC. www.hhmi.org/grants/pdf/08-209_AAMC-HHMI_report.pdf (accessed 22 May 2013).

Baum DA, Smith S, Donovan SSS (2005). The tree-thinking challenge. Science *310*, 979–980.

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

Bowling BV, Acra EE, Wang L, Myers MF, Dean GE, Markle GC, Moskalik CL, Huether CA (2008). Development and evaluation of a genetics literacy assessment instrument for undergraduates. Genetics *178*, 15–22.

Bridgeman B, Morgan R (1996). Success in college for students with discrepancies between performance on multiple-choice and essay tests. J Educ Psychol *88*, 333.

Campbell NA, Reece JB, Urry LA, Cain ML, Wasserman SA, Minorsky PV, Jackson RB (2008). Biology, 8th ed., San Francisco, CA: Pearson/Benjamin Cummings.

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

Chi MTH (2006b). Two approaches to the study of experts' characteristics. In: Cambridge Handbook of Expertise and Expert Performance, ed. KA Ericsson, N Charness, P Feltovich, and R Hoffman, New York: Cambridge University Press, 121–30.

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 14 October 2013).

College Board (2013). AP Biology Curriculum Framework. http://media.collegeboard.com/digitalServices/pdf/ap/10b_2727_AP_Biology_CF_WEB_110128.pdf (accessed 22 May 2013).

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

Dreyfus H, Dreyfus S (2005). Expertise in real world contexts. Organ Stud *26*, 779–792.

Fisher KM, Williams KS, Lineback J (2011). Osmosis and diffusion conceptual assessment. CBE Life Sci Educ *10*, 418–429.

Hartley LM, Wilke BJ, Schramm JW, D'Avanzo C, Anderson CW (2011). College students' understanding of the carbon cycle: contrasting principle-based and informal reasoning. BioScience *61*, 65–75.

Hickman C, Roberts L, Keen S, Larson A, l'Anson H, Eisenhour D (2007). Integrated Principles of Zoology, 14th ed., New York: McGraw-Hill.

IBM (2011). IBM SPSS Statistics for Macintosh, Version 20.0, Armonk, NY: IBM.

JMP (2010). JMP, Version 9, Cary, NC: SAS Institute, 1989–2012.

Klymkowsky MW, Garvin-Doxas K, Zeilik M (2003). Bioliteracy and teaching efficacy: what biologists can learn from physicists. Cell Biol Educ *2*, 155–161.

Kuhn HW (1955). The Hungarian method for the assignment problem. Nav Res Log Q *2*, 83–97.

Labov JB, Reid AH, Yamamoto KR (2010). Integrated biology and undergraduate science education: a new biology education for the twenty-first century? CBE Life Sci Educ *9*, 10–16.

Marbach-Ad G *et al.* (2009). Assessing student understanding of host pathogen interactions using a concept inventory. J Microbiol Educ *10*, 43–50.

Mason A, Singh C (2011). Assessing expertise in introductory physics using categorization task. Phys Rev ST Phys Educ Res *7*, 020110-1–020110-17.

Nehm R, Schonfeld IS (2008). Measuring knowledge of natural selection: a comparison of the CINS, an open-response instrument, and an oral interview. J Res Sci Teach *45*, 1131–1160.

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

Odom AL, Barrow LH (1995). The development and application of a two-tiered diagnostic test measuring college biology students' understanding of diffusion and osmosis following a course of instruction. J Res Sci Teach *32*, 45–61.

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

Rabinowitz HK, Mohammadreza H (1989). A comparison of the modified essay question and multiple choice question formats: their relationship to clinical performance. Family Med *21*, 364.

Raven PH, Johnson GB, Mason KA, Losos J, Singer SR (2011). Biology, 9th ed., New York: McGraw-Hill.

Semsar K, Knight JK, Birol G, Smith MK (2011). The Colorado Learning Attitudes about Science Survey (CLASS) for use in Biology. CBE Life Sci Educ *10*, 268–278.

Shi J, Wood WB, Martin JM, Guild NA, Vicens Q, Knight JK (2010). Diagnostic assessment for introductory molecular and cell biology. CBE Life Sci Educ *9*, 453–461.

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

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

Udovic D, Morris D, Dickman A, Postlethwait J, Wetherwax P (1996). The Workshop Biology Curriculum Handbook, Eugene: University of Oregon.

Wilson CD, Anderson CW, Heidemann M, Merrill JE, Merritt BW, Richmond G, Sibley DF, Parker JM (2007). Assessing students' ability to trace matter in dynamic systems in cell biology. CBE Life Sci Educ *5*, 323–331.

Wood WB (2009). Revising the AP biology curriculum. Science *325*, 1627–1628.

Woodin T, Carter C, Fletcher L (2010). Vision and Change in Biology Undergraduate Education, A Call for Action—initial responses. CBE Life Sci Educ *9*, 71–73.