

Step by Step: Biology Undergraduates' Problem-Solving Procedures during Multiple-Choice Assessment

Luanna B. Prevost[†] and Paula P. Lemons^{†*}

[†]Department of Integrative Biology, University of South Florida, Tampa, FL 33620; [†]Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA 30602

ABSTRACT

This study uses the theoretical framework of domain-specific problem solving to explore the procedures students use to solve multiple-choice problems about biology concepts. We designed several multiple-choice problems and administered them on four exams. We trained students to produce written descriptions of how they solved the problem, and this allowed us to systematically investigate their problem-solving procedures. We identified a range of procedures and organized them as domain general, domain specific, or hybrid. We also identified domain-general and domain-specific errors made by students during problem solving. We found that students use domain-general and hybrid procedures more frequently when solving lower-order problems than higher-order problems, while they use domain-specific procedures more frequently when solving higher-order problems. Additionally, the more domain-specific procedures students used, the higher the likelihood that they would answer the problem correctly, up to five procedures. However, if students used just one domain-general procedure, they were as likely to answer the problem correctly as if they had used two to five domain-general procedures. Our findings provide a categorization scheme and framework for additional research on biology problem solving and suggest several important implications for researchers and instructors.

INTRODUCTION

The call to reform undergraduate education involves shifting the emphasis in science classes away from rote memorization of facts toward learning core concepts and scientific practices (National Research Council [NRC], 2003; American Association for the Advancement of Science [AAAS], 2011). To develop instruction that focuses on core concepts and scientific practices, we need more knowledge about the concepts that are challenging for students to learn. For example, biology education research has established that students struggle with the concepts of carbon cycling (e.g., Anderson *et al.*, 1990; Hartley *et al.*, 2011) and natural selection (e.g., Nehm and Reilly, 2007), but we know much less about students' conceptual difficulties in ecology and physiology. Researchers and practitioners also need to discover how students develop the ability to use scientific practices. Although these efforts are underway (e.g., Anderson *et al.*, 2012; Gormally *et al.*, 2012; Dirks *et al.*, 2013; Brownell *et al.*, 2014), many research questions remain. As research accumulates, educators can create curricula and assessments that improve student learning for all. We investigate one key scientific practice that is understudied in biology education, problem solving (AAAS, 2011; Singer *et al.*, 2012).

For the purposes of this article, we define problem solving as a decision-making process wherein a person is presented with a task, and the path to solving the task is uncertain. We define a problem as a task that presents a challenge that cannot be solved automatically (Martinez, 1998). Problem-solving research began in the 1940s and 1950s and focused on problem-solving approaches that could be used to solve any

Hannah Sevian, Monitoring Editor

Submitted December 17, 2015; Revised June 10, 2016; Accepted June 11, 2016

CBE Life Sci Educ December 1, 2016 15:ar71

DOI:10.1187/cbe.15-12-0255

*Address correspondence to: Paula P. Lemons (plemons@uga.edu).

© 2016 L. B. Prevost and P. P. Lemons. CBE—Life Sciences Education © 2016 The American Society for Cell Biology. This article is distributed by The American Society for Cell Biology under license from the author(s). It is available to the public under an Attribution–Noncommercial–Share Alike 3.0 Unported Creative Commons License (<http://creativecommons.org/licenses/by-nc-sa/3.0>).

"ASCB®" and "The American Society for Cell Biology®" are registered trademarks of The American Society for Cell Biology.

problem regardless of the discipline (Duncker and Lees, 1945; Polya, 1957; Newell and Simon, 1972; Jonassen, 2000, 2012; Bassok and Novick, 2012). Despite the broad applicability of these domain-general problem-solving approaches, subsequent research has shown that the strongest problem-solving approaches derive from deep knowledge of a domain (Newell and Simon, 1972; Chi *et al.*, 1981; Pressley *et al.*, 1987). Domain is a term that refers to a body of knowledge that can be broad, like biology, or narrow, like ecosystem structure and function. This body of literature has developed into a theoretical framework called domain-specific problem solving. We situate our research within this theoretical framework.

THE THEORETICAL FRAMEWORK OF DOMAIN-SPECIFIC PROBLEM SOLVING

Domain-specific problem solving has its origins in information-processing theory (IPT; Newell and Simon, 1972). IPT focuses on the cognitive processes used to reach a problem solution and emphasizes the general thinking processes people use when they attempt problem solving, such as brainstorming (Runco and Chand, 1995; Halpern, 1997) and working backward by beginning with the problem goal and working in reverse toward the initial problem state (Newell *et al.*, 1958; Chi and Glaser, 1985). Despite the empirical evidence for general thinking processes, one of IPT's shortcomings as a comprehensive view of human cognition (Dawson, 1998) is that the knowledge base of the problem solver is not considered.

Domain-specific problem solving expands IPT to recognize that experts in a particular domain have a relatively complete and well-organized knowledge base that enables them to solve the complex problems they face (e.g., Chase and Simon, 1973). One of the landmark studies showing the differences between the knowledge base of experts and nonexperts, or novices, was conducted in science, specifically in physics. Chi and colleagues (1981) compared the classification of physics problems by advanced physics PhD students (i.e., experts) and undergraduates who had just completed a semester of mechanics (i.e., novices), identifying fundamental differences. Chemistry researchers built on Chi's work to identify differences in how experts and novices track their problem solving and use problem categorization and multiple representations (Bunce *et al.*, 1991; Kohl and Finkelstein, 2008; Catrette and Bodner, 2010). Biology researchers built upon this work by conducting similar problem-solving studies among experts and novices in evolution and genetics (Smith, 1992; Smith *et al.*, 2013; Nehm and Ridgway, 2011). Taken together, these studies established that experts tend to classify problems based on deep, conceptual features, while novices classify problems based on superficial features that are irrelevant to the solution.

Domain-specific problem-solving research within biology also has revealed important individual differences within groups of problem solvers. These studies show that wide variation in problem-solving performance exists. For example, some novices who solve problems about evolution classify problems and generate solutions that are expert-like, while others do not (Nehm and Ridgway, 2011). This research points to the importance of studying variations in problem solving within novice populations.

Given the centrality of the knowledge base for domain-specific problem solving, it is necessary to describe the components of

that knowledge base. Domain-specific problem-solving research recognizes three types of knowledge that contribute to expertise. Declarative knowledge consists of the facts and concepts about the domain. Procedural knowledge represents the how-to knowledge that is required to carry out domain-specific tasks. Conditional knowledge describes the understanding of when and where to use one's declarative and procedural knowledge (Alexander and Judy, 1988). Note that the field of metacognition also uses this three-type structure to describe metacognitive knowledge, or what you know about your own thinking (Brown, 1978; Jacobs and Paris, 1987; Schraw and Moshman, 1995). However, for this paper, we use these terms to describe knowledge of biology, not metacognitive knowledge. More specifically, we focus on procedural knowledge.

Procedural knowledge consists of procedures. Procedures are tasks that are carried out automatically or intentionally during problem solving (Alexander and Judy, 1988). Procedures exist on a continuum. They can be highly specific to the domain, such as analyzing the evolutionary relationships represented by a phylogenetic tree, or general and applicable to problems across many domains, such as paraphrasing a problem-solving prompt (Pressley *et al.*, 1987, 1989; Alexander and Judy, 1988).

APPLYING DOMAIN-SPECIFIC PROBLEM SOLVING TO MULTIPLE-CHOICE ASSESSMENT IN BIOLOGY

We used domain-specific problem solving to investigate the most common form of assessment in the college biology classroom, multiple-choice assessment (Zheng *et al.*, 2008; Momsen *et al.*, 2013). College biology and science, technology, engineering, and mathematics (STEM) courses rely on multiple-choice assessment due to large enrollments, limited teaching assistant support, and ease of scoring. Outside the classroom, multiple-choice assessment is used on high-stakes exams that determine acceptance to professional schools, like the Medical College Admissions Test and Graduate Record Exam. To our knowledge, the framework of domain-specific problem solving has not been applied previously to investigate multiple-choice assessment in college biology.

It has become common practice within the biology education community to think about assessment, including multiple-choice assessment, by determining the Bloom's taxonomy ranking of assessment items (e.g., Bissell and Lemons, 2006; Crowe *et al.*, 2008; Momsen *et al.*, 2010, 2013). Bloom's *Taxonomy of Educational Objectives* was built to facilitate the exchange of test items among faculty; it was not based primarily on the evaluation of student work (Bloom, 1956; Anderson and Krathwohl, 2001). Bloom's taxonomy helps educators think about the range of cognitive processes they could ask their students to perform and has served as an invaluable resource enabling educators to improve alignment between learning objectives, assessments, and classroom curricula (e.g., Crowe *et al.*, 2008). When applying Bloom's taxonomy to assessment items, items are ranked as remembering, understanding, applying, analyzing, evaluating, and synthesizing. Items ranked as remembering and understanding are grouped as lower-order items; and items ranked as applying, analyzing, evaluating, and synthesizing are grouped as higher-order items (Zoller, 1993; Crowe *et al.*, 2008). Despite the value of Bloom's taxonomy for instructors, what is not known is the relationship

between the procedural knowledge of domain-specific problem solving and the Bloom's ranking of biology assessments. This is a critical gap in the literature, because efforts to improve student learning in college science classrooms may be stymied if critical insights about student work from domain-specific problem solving are not linked to our understanding of assessment and curricular design.

In the study reported here, we used the theoretical lens of domain-specific problem solving to describe the procedural knowledge of nonmajors in an introductory biology course. We addressed the following research questions:

1. What are the domain-general and domain-specific procedures students use to solve multiple-choice biology problems?
2. To what extent do students use domain-general and domain-specific procedures when solving lower-order versus higher-order problems?
3. To what extent does the use of domain-general or domain-specific procedures influence the probability of answering problems correctly?

METHODS

Setting and Participants

We recruited participants from a nonmajors introductory biology course at a southeastern public research university in the Spring 2011 semester. One of the authors (P.P.L.) was the course instructor. The course covered four major areas in biology: evolution, ecology, physiology, and organismal diversity. The instructor delivered course content using lecture interspersed with clicker questions and additional opportunities for students to write and discuss. Students also completed five in-class case studies during the semester; students completed cases in self-selected small groups and turned in one completed case study per group for grading. In addition to group case studies, the instructor assessed student learning via individual exams. Students also received points toward their final grades based on clicker participation.

In the second week of the semester, the instructor announced this research study in class and via the course-management system, inviting all students to participate. Students who volunteered to participate by completing an informed consent form were asked to produce written think-alouds for problems on course exams throughout the semester. One hundred sixty-four students completed an informed consent form. Of the 164 consenting students, 140 students actually produced a written think-aloud for at least one of 13 problems; of the 140 students, 18 did written think-alouds for all 13 problems. The remainder of students did written think-alouds for one to 13 problems. On average, research participants provided written think-alouds for 7.76 problems.

The 164 consenting students represented 73.9% of the course enrollment ($n = 222$). The 164 consenting students included 70.8% females and 29.2% males; 20.4% freshmen, 40.9% sophomores, 24.1% juniors, and 13.9% seniors. The 164 students were majoring in the following areas: 3.7% business, 1.5% education, 4.4% humanities, 11.0% life and physical sciences, 5.9% engineering, and 72.3% social sciences.

This research was conducted under exempt status at the University of Georgia (UGA; IRB project 201110340).

Data Collection

Problem Development. We wrote 16 multiple-choice problems to include in this study. All problems related to material dealt with during class and focused specifically on ecosystems, evolution, and structure–function relationships. On data analysis, three problems were excluded, because most students were confused by the wording or visual representations or were able to solve the problem correctly with a superficial strategy. Each problem was preceded by a prompt for students to provide their written think-aloud (see *Written Think-Alouds* section). Each problem was also labeled with a preliminary Bloom's taxonomy categorization (Anderson and Krathwohl, 2001). A summary of all problems, including a description, the preliminary Bloom's ranking, and the faculty consensus Bloom's ranking, is provided in Table 1. As an example, one of the final 13 problems is shown in Figure 1. All other problems are shown in Supplemental Figure S1.

Ranking of Problems by Bloom's Level. We wanted to investigate the use of domain-general or domain-specific procedures in lower-order versus higher-order problems. We asked three biology faculty members who were not investigators in this study to rank the Bloom's levels of the problems we developed. The biology faculty members were selected because they have extensive teaching experience in college biology and also have experience ranking assessment items using Bloom's taxonomy. The faculty used a protocol similar to one described previously (Momsen *et al.*, 2010). To assist with Bloom's ranking, we provided them with class materials relevant to the problems, including lecture notes and background readings. This is necessary, because the ranking of a problem depends on the material that students have encountered in class previously. The faculty members independently ranked each problem. Interrater reliability of independent rankings was determined using an intra-class coefficient (0.82). The faculty members met to discuss their rankings and settled disagreements by consensus. The preliminary Bloom's rankings and the faculty consensus Bloom's rankings for problems are reported in Table 1. For the remainder of the paper, we use the consensus Bloom's rankings to describe problems as either lower order or higher order.

Administration of Problems to Students. The 13 problems included in this study were administered to students on exams 1, 2, 3, and the final exam as follows: three on exam 1, three on exam 2 four on exam 3, and three on the final exam. Students' multiple-choice responses were part of the actual exam score. They received 0.5 extra-credit points for providing satisfactory documentation of their thought processes. Students did not receive extra credit if we judged their documentation to be insufficient. Insufficient responses were those in which students made only one or two brief statements about their problem-solving process (e.g., "I chose C"). Students could answer the multiple-choice problem and opt not to provide documentation of their thinking for extra credit. Students could receive up to 6.5 points of extra credit for documentation of the problem set. The total points possible for the semester were 500, so extra credit for this research could account for up to 1.3% of a student's grade.

Written Think-Alouds. We developed a protocol to capture students' written descriptions of their thought processes while

TABLE 1. Summary of problems used for data collection

Problem number	Exam number	Description (*indicates problems with visual representations)	Preliminary Bloom's ranking	Faculty consensus Bloom's ranking
1	1	*Evolution problem asking students to choose an explanation that best describes the phylogenetic relationships presented	Understanding	Understanding
2	1	Ecology problem asking students to choose an example of resource partitioning	Applying	Understanding
3	1	*Evolution problem asking students to choose the best conclusion about species relationships between stickleback populations based on morphological and population data	Applying and Analyzing	Analyzing
4	2	* <i>Echinacea</i> clinical trials problem asking students to choose the effect of treatment with <i>Echinacea</i> on upper respiratory symptoms	Applying	Understanding
5	2	Animal evolution problem asking students to choose which pieces of evidence support a hypothesis	Understanding	Recalling
6	2	*Human evolution problem asking students to choose the best placement of <i>Australopithecus afarensis</i> on a phylogenetic tree	Understanding	Applying
7	3	Blood cell structure/function problem asking students to choose which descriptions exemplify structure matching function	Understanding	Recalling
8	3	*Neuron structure–function problem asking students to choose the correct status of voltage-gated channels (open or closed) based on oscilloscope data	Understanding	Understanding
9	3	*Neuron structure–function problem asking students to choose the best hypothesis to explain oscilloscope data from a neurotoxin experiment	Applying and Analyzing	Analyzing
10	3	*Mammalian structure–function and evolution problem asking students to choose the likely geographic location of three hypothetical mammals based on their morphology	Applying and Analyzing	Applying
11	Final	*Human population problem asking students to choose the correct description of the trend in human population growth based on the annual rate of increase	Inadvertently not ranked	Recalling
12	Final	*Ecosystem ecology problem asking students to choose the observations that are most likely to be made before and after the introduction of a predator to the ecosystem	Applying	Applying
13	Final	*Evolution problem asking students to choose a graph that best predicts the amount of krait venom required to kill eels in populations of eels that exist with and without krait	Applying and Analyzing	Analyzing

For each problem, a description is included along with the preliminary Bloom's ranking, and the final consensus Bloom's ranking. The actual problems are included in Supplemental Figure S1.



solving problems on exams based on a think-aloud interview approach. In the think-aloud interview approach, research participants are given a problem to solve and are asked to say aloud everything they are thinking while solving the problem (Ericsson and Simon, 1984; Keys, 2000). In the written think-aloud, students are asked to write, rather than say aloud, what they are thinking as they solve a problem. To train students to perform a written think-aloud, the course instructor modeled the think-aloud in class. She then assigned a homework problem that required students to answer a multiple-choice problem and construct written think-alouds recounting how they solved the problem. We then reviewed students' homework and provided feedback. We selected examples of good documentation and poor documentation and published these anonymously on the

online course-management system. After this training and feedback, we included four problems on every exam for which we asked students to provide a written think-aloud description. We collected 1087 written think-alouds from 140 students (63% of course enrollment, $n = 222$) for 13 problems. Figure 2 shows a typical example of a student written think-aloud.

Data Analysis

We analyzed students' written think-alouds using a combination of qualitative and quantitative methods. We used qualitative content analysis (Patton, 1990) to identify and categorize the primary patterns of student thinking during problem solving. We used quantitative analysis to determine the relationship between use of domain-general, hybrid, and domain-specific

Sticklebacks are small fish found in a variety of habitats in the Northern Hemisphere. Two forms of sticklebacks have been identified – the Benthics and the Limnetics. Examine the pictures of the two stickleback forms and the data in Table A. What is the best conclusion you can draw from these data?

Benthic male Limnetic male

Table A. Proportions of benthics, limnetics, and hybrids found in traps. For three different years, traps were set for fish in Paxton Lake in British Columbia. The traps were regularly checked and the type of fish (benthic, limnetic, or benthic/limnetic hybrids) and the number of each type were determined. The relative proportion of each type for a single year is presented with the actual numbers counted in parentheses (McPhail 1992).

Year	Total	Benthics	Limnetics	Hybrids
1	1057	0.50 (528)	0.48 (509)	0.019 (20)
2	982	0.50 (479)	0.49 (473)	0.010 (10)
3	994	0.49 (491)	0.49 (489)	0.014 (14)

A. Based on the morphological species concept Benthic and Limnetic fish are different species.
 B. Based on the phylogenetic species concept Benthic and Limnetic fish are the same species.
 C. Based on the biological species concept Benthic and Limnetic fish are the same species.
 D. Based on the biological species concept Benthic and Limnetic fish are different species.
 E. I need more information to draw a conclusion about whether Benthic and Limnetics are the same or different species.

FIGURE 1. Sample problem from the domain of evolution used to probe students' problem-solving procedures. The preliminary ranking that students saw for this question was Applying and Analyzing based on Bloom's taxonomy. Experts ranked this problem as Analyzing. The correct answer is E. Images of benthic and limnetic males are courtesy of Elizabeth Carefoot, Simon Fraser University.

procedures and problem type and to investigate the impact of domain-general/hybrid and domain-specific procedure use on answering correctly.

Qualitative Analyses of Students' Written Think-alouds. The goal of our qualitative analysis was to identify the cognitive procedures students follow to solve multiple-choice biology problems during an exam. Our qualitative analysis took place in two phases.

Phase 1: Establishing Categories of Student Problem-Solving Procedures. Independently, we read dozens of individual think-alouds for each problem. While we read, we made notes about the types of procedures we observed. One author (P.P.L.) noted, for example, that students recalled concepts,

organized their thinking, read and ruled out multiple-choice options, explained their selections, and weighed the pros and cons of multiple-choice options. The other author (L.B.P.) noted that students recalled theories, interpreted a phylogenetic tree, identified incomplete information, and refuted incorrect information. After independently reviewing the written think-alouds, we met to discuss what we had found and to build an initial list of categories of problem-solving procedures. Based on our discussion, we built a master list of categories of procedures (Supplemental Table S1).

Next, we compared our list with Bloom's *Taxonomy of Educational Objectives* (Anderson and Krathwohl, 2001) and the *Blooming Biology Tool* (Crowe *et al.*, 2008). We sought to determine whether the cognitive processes described in these sources corresponded to the cognitive processes we observed in

1. Read the question.
2. Know I'm looking for answer that includes viable fertile offspring. The fish have some similarities but not enough difference that I can't judge if they are the same species by sight. I also don't know about habitat or diet variances.
3. Looked over the chart.
4. The chart shows that hybrids can be produced but gives no information if they are viable and fertile.
5. Notice that just because number of hybrids increased from year 2 to 3 (10 to 14 hybrids) does not mean they can reproduce. It could mean that the benthics and limnetics produced more hybrids that year.
6. I don't have enough information to solve the problems so I choose answer E.

FIGURE 2. Written think-aloud from an introductory biology student who had been instructed to write down her procedures for solving a multiple-choice biology problem. This document describes the student's procedures for solving the problem shown in Figure 1.

our initial review of students' written think-alouds. Where there was overlap, we renamed our categories to use the language of Bloom's taxonomy. For the categories that did not overlap, we kept our original names.

Phase 2: Assigning Student Problem-Solving Procedures to Categories. Using the list of categories developed in phase 1, we categorized every problem-solving procedure articulated by students in the written think-alouds. We analyzed 1087 documents for 13 problems. For each of the 13 problems, we followed the same categorization process. In a one-on-one meeting, we discussed a few written think-alouds. While still in the same room, we categorized several written think-alouds independently. We then compared our categorizations and discussed any disagreements. We then repeated these steps for additional think-alouds while still together. Once we reached agreement on all categories for a single problem, we independently categorized a common subset of written think-alouds to determine interrater reliability. When interrater reliability was below a level we considered acceptable (0.8 Cronbach's alpha), we went through the process again. Then one author (either L.B.P. or P.P.L.) categorized the remainder of the written think-alouds for that problem.

At the end of phase 2, after we had categorized all 1087 written think-alouds, we refined our category list, removing categories with extremely low frequencies and grouping closely related categories. For example, we combined the category Executing with Implementing into a category called Analyzing Visual Representations.

Phase 3: Aligning Categories with Our Theoretical Framework. Having assigned student problem-solving procedures to categories, we determined whether the category aligned best with domain-general or domain-specific problem solving. To make this determination, we considered the extent to which the problem-solving procedures in a category depended on knowledge of biology. Categories of procedures aligned with domain-general problem solving were carried out without drawing on content knowledge (e.g., Clarifying). Categories aligned with domain-specific problem solving were carried out using content knowledge (e.g., Checking). We also identified two categories of problem solving that we labeled hybrids of domain-general and domain-specific problem solving, because students used content knowledge in these steps, but they did so superficially (e.g., Recognizing).

Supplemental Table S1 shows the categories that resulted from our analytical process, including phase 1 notes, phase 2 categories, and phase 3 final category names as presented in this paper. Categories are organized into the themes of domain-general, hybrid, and domain-specific problem solving (Supplemental Table S1).

Quantitative Analyses of Students' Written Think-Alouds. To determine whether students used domain-general/hybrid or domain-specific problem solving preferentially when solving problems ranked by faculty as lower order or higher order, we used generalized linear mixed models (GLMM). GLMM are similar to ordinary linear regressions but take into account nonnormal distributions. GLMM can also be applied to unbalanced repeated measures (Fitzmaurice *et al.*, 2011). In our data set, an individual student could provide documentation to one or more

problems (up to 13 problems). Thus, in some but not all cases, we have repeated measures for individuals. To account for these repeated measures, we used "student" as our random factor. We used the problem type (lower order or higher order) as our fixed factor. Because our independent variables, number of domain-general/hybrid procedures and number of domain-specific procedures, are counts, we used a negative binomial regression. For this analysis and subsequent quantitative analyses, we grouped domain-general and hybrid procedures. Even though hybrid procedures involve some use of content knowledge, the content knowledge is used superficially; we specifically wanted to investigate the impact of weak content-knowledge use compared with strong content-knowledge use. Additionally, the number of hybrid procedures in our data set is relatively low compared with domain-general and domain-specific.

To determine whether students who used more domain-general/hybrid procedures or domain-specific procedures were more likely to have correct answers to the problems, we also used GLMM. We used the number of domain-general/hybrid procedures and the number of domain-specific procedures as our fixed factors and student as our random factor. In this analysis, our dependent variable (correct or incorrect response) was dichotomous, so we used a logistic regression (Fitzmaurice *et al.*, 2011). We also explored the correlations between the average number of domain-general/hybrid and domain-specific procedures used by students and their final percentage of points for the course.

RESULTS

In this section, we present the results of our analyses of students' procedures while solving 13 multiple-choice, biology problems (Figure 1 and Supplemental Figure S1). We used the written think-aloud protocol to discover students' problem-solving procedures for all 13 problems.

Students Use Domain-General and Domain-Specific Procedures to Solve Multiple-Choice Biology Problems

We identified several categories of procedures practiced by students during problem solving, and we organized these categories based on the extent to which they drew upon knowledge of biology. Domain-general procedures do not depend on biology content knowledge. These procedures also could be used in other domains. Hybrid procedures show students assessing multiple-choice options with limited and superficial references to biology content knowledge. Domain-specific procedures depend on biology content knowledge and reveal students' retrieval and processing of correct ideas about biology.

Domain-General Procedures. We identified five domain-general problem-solving procedures that students practiced (Table 2). Three of these have been described in Bloom's taxonomy (Anderson and Krathwohl, 2001). These include Analyzing Domain-General Visual Representations, Clarifying, and Comparing Language of Options. In addition, we discovered two other procedures, Correcting and Delaying, that we also categorized as domain general (Table 2).

During Correcting, students practiced metacognition. Broadly defined, metacognition occurs when someone knows, is aware of, or monitors his or her own learning (White, 1998). When students corrected, they identified incorrect thinking

TABLE 2. Students' problem-solving procedures while solving multiple-choice biology problems

Problem-solving procedures	Description: this category refers to parts of the written think-aloud in which students ...
Domain-general procedures	
Analyzing Domain-General Visual Representations ^{a,b}	For a visual representation that is not unique to biology (e.g., a table or a bar graph), broke it down and determined how the individual parts related to one another.
Clarifying ^a	Restated or paraphrased the problem stem or one of the multiple-choice options.
Comparing Language of Options ^a	Detected similarities and differences in the language of two multiple-choice options.
Correcting ^b	Pointed out that they had been thinking incorrectly about the problem earlier in the written think-aloud and now see the correct way to think about the problem.
Delaying ^b	Considered one of the multiple-choice options and decided that it should not be eliminated. Rather, the quality of that option should be evaluated later, after the other multiple-choice options are considered.
Hybrid procedures	
Comparing Correctness of Options ^a	Detected similarities and differences in two multiple-choice options, often based on a superficial evaluation of the content of the options (e.g., one option appears more correct than another).
Recognizing ^a	Noted that a multiple-choice option is correct or incorrect without any rationale.
Domain-specific procedures	
Adding Information ^b	Provided more information about one of the multiple-choice options, such as additional facts that were omitted or corrections to incorrect statements (i.e., presented incorrectly to serve as distractors).
Analyzing Domain-Specific Visual Representation ^{a,b}	For a visual representation that is unique to biology (e.g., a phylogenetic tree or food web), broke it down and determined how the individual parts related to one another.
Asking a Question ^c	Asked a question about the problem stem or multiple-choice options.
Checking ^a	Explained why an option is correct or incorrect by comparing the option with their knowledge or with the data provided in the problem.
Predicting ^{a,c}	As an early step in the written think-aloud, predicted what they expected the answer to be (i.e., what multiple-choice option they were looking for).
Recalling ^a	Retrieved basic facts or concepts from class, notes, or the textbook (i.e., declarative knowledge).

The procedures are categorized as domain-general, hybrid, and domain-specific. Superscripts indicate whether the problem-solving procedure aligns with previously published conceptions of student thinking or was newly identified in this study: ^aAnderson and Krathwohl (2001); ^bidentified in this study; ^cCrowe *et al.* (2008).

they had displayed earlier in their written think-aloud and mentioned the correct way of thinking about the problem.

When students Delayed, they described their decision to postpone full consideration of one multiple-choice option until they considered other multiple-choice options. We interpreted these decisions as students either not remembering how the option connected with the question or not being able to connect that option to the question well enough to decide whether it could be the right answer.

Hybrid Procedures. We identified two problem-solving procedures that we categorized as hybrid, Comparing Correctness of Options and Recognizing. Students who compared correctness of options stated that one choice appeared more correct than the other without giving content-supported reasoning for their choice. Similarly, students who recognized an option as correct did not support this conclusion with a content-based rationale.

Domain-Specific Procedures. In our data set, we identified six domain-specific problem-solving procedures practiced by students (Table 2). Four of these have been previously described. Specifically, Analyzing Domain-Specific Visual Representations, Checking, and Recalling were described in Bloom's taxonomy (Anderson and Krathwohl, 2001). Predicting was described by Crowe and colleagues (2008). We identified two additional categories of domain-specific problem-solving procedures practiced by students who completed our problem set, Adding Information and Asking a Question.

Adding Information occurred when students recalled material that was pertinent to one of the multiple-choice options and incorporated that information into their explanations of why a particular option was wrong or right.

Asking a Question provides another illustration of students practicing metacognition. When students asked a question, they pointed out that they needed to know some specific piece of content that they did not know yet. Typically, students who asked a question did so repeatedly in a single written think-aloud.

Students Make Errors While Solving Multiple-Choice Biology Problems

In addition to identifying domain-general, hybrid, and domain-general procedures that supported students' problem-solving, we identified errors in students' problem solving. We observed six categories of errors, including four that we categorized as domain general and two categorized as domain specific (Table 3).

The domain-general errors include Contradicting, Disregarding Evidence, Misreading, and Opinion-Based Judgment. In some cases, students made statements that they later contradicted; we called this Contradicting. Disregarding Evidence occurred when students' failed to indicate use of evidence. Several problems included data in the question prompt or in visual representations. These data could be used to help students select the best multiple-choice option, yet many students gave no indication that they considered these data. When students' words led us to believe that they

TABLE 3. Students' errors while solving multiple-choice biology problems

Problem-solving errors	Description: this category refers to parts of the written think-aloud in which the student ...	Example quotes
Domain-general errors		
Contradicting	Stated two ideas that were in opposition to each other.	(C) says that are they same based on the biological species concept. The data that proves there are hybrids proves this to be true. I mark it. (E) could also make sense but I think there is enough information to make a decision.
Disregarding Evidence	Did not use some or all of the data provided in the problem.	A. Incorrect answer—the data does not represent morphological characteristics, so cannot conclude this answer. Move on.
Misreading	Read the question prompt or answer options incorrectly	B is incorrect because Atlantic eels should show some resistance since the Atlantic eel have developed in the presence of krait toxin.
Opinion-Based Judgment	Gave an opinion and did not use biology content knowledge.	E may be right, but I feel confident with C.
Domain-specific errors		
Making Incorrect Assumptions	Stated that the graph or other visual representation provides no useful information.	Examine graph. Hybrids are not seeming to live (not viable).
Misunderstanding Content	Showed incorrect understanding of content knowledge.	(C) says that are they same based on the biological species concept. The data that proves there are hybrids proves this to be true. I mark it.

The errors are presented in alphabetical order, described, and illustrated with example quotes from different students' documentation of their solutions to the problem shown in Figure 1 (except for Misreading, which is from problem 13 in Supplemental Figure S1).

did not examine the data, we assigned the category Disregarding Evidence. Students also misread the prompt or the multiple-choice options, and we termed this Misreading. For example, Table 3 shows the student Misreading; the student states that Atlantic eels are in the presence of krait toxins, whereas the question prompt stated there are no krait in the Atlantic Ocean. In other cases, students stated that they arrived at a decision based on a feeling or because that option just seemed right. For example, in selecting option C for the stickleback problem (Figure 1), one student said, "E may be right, but I feel confident with C. I chose Answer C." These procedures were coded as Opinion-Based Judgment.

We identified two additional errors that we classified as domain specific, Making Incorrect Assumptions and Misunderstanding Content. Making Incorrect Assumptions was identified when students made faulty assumptions about the information provided in the prompt. In these cases, students demonstrated in one part of their written think-aloud that they understood the conditions for or components of a concept. However, in another part of the written think-aloud, students assumed the presence or absence of these conditions or components without carefully examining whether they held for the given problem. In the example shown in Table 3, the student assumed additional information on fertility that was not provided in the problem. We classified errors that showed a poor understanding of the biology content as Misunderstanding Content. Misunderstanding Content was exhibited when students stated incorrect facts from their long-term memory, made false connections between the material presented and biology concepts, or showed gaps in their understanding of a concept. In the Misunderstanding Content example shown in Table 3, the student did not understand that the biological species concept requires two conditions, that is, the offspring must be viable and fertile. The student selected the biological species

concept based only on evidence of viability, demonstrating misunderstanding.

To illustrate the problem-solving procedures described above, we present three student written think-alouds (Table 4, A–C). All three think-alouds were generated in response to the stickleback problem; pseudonyms are used to protect students' identities (Figure 1). Emily correctly solved the stickleback problem using a combination of domain-general and domain-specific procedures (Table 4A). She started by thinking about the type of answer she was looking for (Predicting). Then she analyzed the stickleback drawings and population table (Analyzing Domain-General Visual Representations) and explained why options were incorrect or correct based on her knowledge of species concepts (Checking). Brian (Table 4B) took an approach that included domain-general and hybrid procedures. He also made some domain-general and domain-specific errors, which resulted in an incorrect answer; Brian analyzed some of the domain-general visual representations presented in the problem but disregarded others. He misunderstood the content, incorrectly accepting the biological species concept. He also demonstrated Recognizing when he correctly eliminated choice B without giving a rationale for this step. In our third example (Table 4C), Jessica used domain-general, hybrid, and domain-specific procedures, along with a domain-specific error, and arrived at an incorrect answer.

Domain-Specific Procedures Are Used More Frequently for Higher-Order Problems Than Lower-Order Problems

To determine the extent to which students use domain-general and domain-specific procedures when solving lower-order versus higher-order problems, we determined the frequency of domain-general and hybrid procedures and domain-specific procedures for problems categorized by experts as lower order or higher order. We grouped

TABLE 4. Students' written think-alouds describing their processes for solving the stickleback problem

Part A	
Problem-solving procedures	Written think-aloud—Emily
PREDICTING	Read the question.
analyzing domain-general visual representations,	Know I'm looking for answer that includes viable fertile offspring.
CHECKING	The fish have some similarities, but enough differences that I can't judge if they are the same species by sight. I also don't know about habitat or diet variances.
analyzing domain-general visual representations	Looked over the chart.
	The chart shows that the hybrids can be produced but gives no information if they are viable and fertile.
ADDING INFORMATION, CHECKING	Noticed that just because number of hybrids increased from year 2 to 3 (10–14 hybrids) does not mean they can reproduce. It could mean that the benthics and limnetics produced more hybrids that year.
CHECKING	I don't have enough information to solve this problem, so I choose answer E.
Part B	
Problem-solving procedures	Written think-aloud—Brian
analyzing domain-general visual representations	Read question twice.
MISUNDERSTANDING CONTENT	Look at chart. Notice that they can interbreed, meaning they are one species according to the biological species concept.
disregarding evidence	Read A. They are the same species—Not A
recognizing	Read B. It isn't the phylogenetic concept being tested—Not B
	Read C. C matches my hypothesis. C is the answer.
Part C	
Problem-solving procedures	Written think-aloud—Jessica
ASKING A QUESTION	Look at info.
RECALLING, MISUNDERSTANDING CONTENT	We're trying to find if they are the same species or not.
	What defines a species?
	BSC → can create viable hybrids that are similar
	Morphologically similar.
	Look at data and answers.
NEGATIVE CHECKING	a. Their shape, etc. isn't described here.
recognizing	b. Don't have their info.
recognizing, delaying	c. Maybe → they did create viable offspring.
recognizing	d. No.
recognizing, delaying	e. Maybe, we only have that they created offspring; not much other info.

Different types of problem-solving processes are indicated with different font types: Domain-general problem-solving steps: blue lowercase font; domain-specific problem-solving steps: blue uppercase font, hybrid problem-solving steps: blue italics; domain-general errors: orange lowercase font; domain-specific errors: orange uppercase font. The written think-alouds are presented in the exact words of the students. A, Emily, all domain-general and domain-specific steps; correct answer: E; B, Brian, domain-general and hybrid steps, domain-general and domain-specific errors; incorrect answer: C; C, Jessica, domain-general, hybrid, and domain-specific steps; domain-specific errors; incorrect answer: C.


domain-general and hybrid procedures, because we specifically wanted to examine the difference between weak and strong content usage. As Table 5, A and B, shows, students frequently used both domain-general/hybrid and domain-specific procedures to solve all problems. For domain-general/hybrid procedures, by far the most frequently used procedure for lower-order problems was Recognizing ($n = 413$); the two most frequently used procedures for higher-order problems were Analyzing Domain-General Representations ($n = 153$) and Recognizing ($n = 105$; Table 5A). For domain-specific procedures, the use of Checking dominated both lower-order ($n = 903$) and higher-order problems ($n = 779$). Recalling also was used relatively frequently for lower-order problems ($n = 207$), as were Analyzing Domain-Specific Visual Representations, Predicting, and Recalling for higher-order problems ($n = 120$, $n = 106$, and $n = 107$, respectively). Overall, students used more domain-general and hybrid procedures when solving

lower-order problems (1.43 ± 1.348 per problem) than when solving higher-order problems (0.74 ± 1.024 per problem; binomial regression $B = 0.566$, $SE = 0.079$, $p < 0.005$). Students used more domain-specific procedures when solving higher-order problems (2.57 ± 1.786 per problem) than when solving lower-order problems (2.38 ± 2.2127 per problem; binomial regression $B = 0.112$, $SE = 0.056$, $p < 0.001$).

Most Problem-Solving Errors Made by Students Involve Misunderstanding Content

We also considered the frequency of problem-solving errors made by students solving lower-order and higher-order problems. As Table 6 shows, most errors were categorized with the domain-specific category Misunderstanding Content, and this occurred with about equal frequency in lower-order and higher-order problems. The other categories of errors were less frequent. Interestingly, the domain-general

TABLE 5. Frequency of each problem-solving procedure for lower-order and higher-order problems



Lower frequencies Higher frequencies

Part A. Domain-general and hybrid procedures								
	Analyzing domain-general visual representations	Clarifying	Comparing language of options	Correcting	Delaying	Comparing correctness of options	Recognizing	Total
Lower-order problems $n = 7$	99	57	129	9	43	18	413	768
Higher-order problems $n = 6$	153	26	28	0	23	5	105	340
Part B. Domain-specific procedures								
	Adding information	Analyzing domain-specific visual representations	Asking a question	Checking	Predicting	Recalling	Total	
Lower-order problems $n = 7$	23	42	95	903	23	207	1293	
Higher-order problems $n = 6$	16	120	37	779	106	107	1165	

Procedures are presented from left to right in alphabetical order. A color scale is used to represent the frequency of each procedure, with the lowest-frequency procedures shown in dark blue, moderate-frequency procedures shown in white, and high-frequency procedures shown in dark red.

errors Contradicting and Opinion-Based Judgment both occurred more frequently with lower-order problems. In contrast, the domain-specific error Making Incorrect Assumptions occurred more frequently with higher-order problems.

Using Multiple Domain-Specific Procedures Increases the Likelihood of Answering a Problem Correctly

To examine the extent to which the use of domain-general or domain-specific procedures influences the probability of answering problems correctly, we performed a logistic regression. Predicted probabilities of answering correctly are shown in Figure 3 for domain-general and hybrid procedures and Figure 4 for domain-specific procedures. Coefficients of the logistic regression analyses are presented in Supplemental Tables S2 and S3. As Figure 3 shows, using zero domain-general or hybrid procedures was associated with a 0.53 predicted probability of being correct. Using one domain-general or hybrid procedure instead of zero increased the predicted probability of correctly answering a problem to 0.79. However, students who used two or more domain-general or hybrid procedures instead of one did not increase the predicted probability of answering a prob-

lem correctly. In contrast, as Figure 4 shows, using zero domain-specific procedures was associated with only a 0.34 predicted probability of answering the problem correctly, and students who used one domain-specific procedure had a 0.54 predicted probability of success. Strikingly, the more domain-specific procedures used by students, the more likely they were to answer a problem correctly up to five procedures; students who used five domain-specific procedures had a 0.97 probability of answering correctly. Predicted probabilities for students using seven and nine domain-specific codes show large confidence intervals around the predictions due to the low sample size ($n = 8$ and 4 , respectively). Also, we examined the extent to which the use of domain-general or domain-specific procedures correlates with course performance. We observed a weak positive correlation between the average number of domain-specific procedures used by students for a problem and their final percentage of points in the course (Spearman's $\rho = 0.306$; $p < 0.001$). There was no correlation between the average number of domain-general/hybrid procedures used by students for a problem and their final percentage of points in the course (Spearman's $\rho = 0.015$; $p = 0.857$).

TABLE 6. Frequency of errors for lower-order and higher-order problems

Lower frequencies Higher frequencies

Part A. Domain-general errors					
	Contradicting	Disregarding Evidence	Misreading	Opinion Based Judgment	Total
Lower-order problems $n = 7$	10	21	10	23	64
Higher-order problems $n = 6$	4	21	12	4	41
Part B. Domain-specific errors					
	Making incorrect assumptions		Misunderstanding content		Total
Lower-order problems $n = 7$	15		210		225
Higher-order problems $n = 6$	30		198		228

Categories of errors are presented from left to right in alphabetical order. A color scale is used to represent the frequency of each type of error, with the lowest-frequency errors shown in dark blue, moderate-frequency errors shown in white, and high-frequency errors shown in dark red.

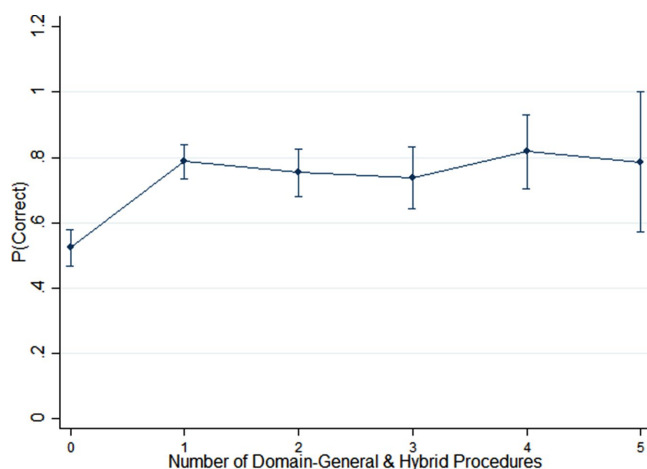


FIGURE 3. Predicted probability of a correct answer based on the number of domain-general and hybrid procedures.

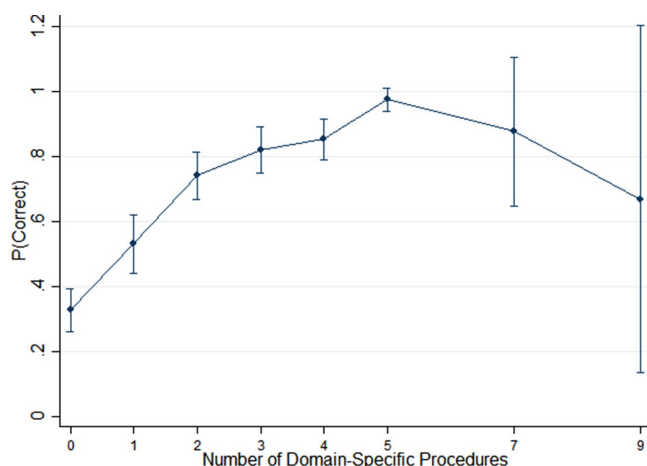


FIGURE 4. Predicted probability of a correct answer based on the number of domain-specific procedures.

DISCUSSION

We have used the theoretical framework of domain-specific problem solving to investigate student cognition during problem solving of multiple-choice biology problems about ecology, evolution, and systems biology. Previously, research exploring undergraduate cognition during problem solving has focused on problem categorization or students' solutions to open-response problems (Smith and Good, 1984; Smith, 1988; Lavoie, 1993; Nehm and Ridgway, 2011; Smith *et al.* 2013). Our goal was to describe students' procedural knowledge, including the errors they made in their procedures. Below we draw several important conclusions from our findings and consider the implications of this research for teaching and learning.

Domain-Specific Problem Solving Should Be Used for Innovative Investigations of Biology Problem Solving

Students in our study used a variety of procedures to solve multiple-choice biology problems, but only a few procedures were used at high frequency, such as Recognizing and Checking. Other procedures that biology educators might most want

students to employ were used relatively infrequently, including Correcting and Predicting. Still other procedures that we expected to find in our data set were all but absent, such as Stating Assumptions. Our research uncovers the range of procedures promoted by multiple-choice assessment in biology. Our research also provides evidence for the notion that multiple-choice assessments are limited in their ability to prompt some of the critical types of thinking used by biologists.

We propose that our categorization scheme and the theoretical framework of domain-specific problem solving should be applied for further study of biology problem solving. Future studies could be done to understand whether different ways of asking students to solve a problem at the same Bloom's level could stimulate students to use different procedures. For example, if the stickleback problem (Figure 1) were instead presented to students as a two-tier multiple-choice problem, as multiple true-false statements, or as a constructed-response problem, how would students' procedures differ? Additionally, it would be useful to investigate whether the more highly desired, but less often observed procedures of Correcting and Predicting are used more frequently in upper-level biology courses and among more advanced biology students.

We also propose research to study the interaction between procedure and content. With our focus on procedural knowledge, we intentionally avoided an analysis of students' declarative knowledge. However, our process of analysis led us to the conclusion that our framework can be expanded for even more fruitful research. For example, one could look within the procedural category Checking to identify the declarative knowledge being accessed. Of all the relevant declarative knowledge for a particular problem, which pieces do students typically access and which pieces are typically overlooked? The answer to this question may tell us that, while students are using an important domain-specific procedure, they struggle to apply a particular piece of declarative knowledge. As another example, one could look within the procedural category Analyzing Visual Representations to identify aspects of the visual representation that confuse or elude students. Findings from this type of research would show us how to modify visual representations for clarity or how to scaffold instruction for improved learning. We are suggesting that future concurrent studies of declarative and procedural knowledge will reveal aspects of student cognition that will stay hidden if these two types of knowledge are studied separately. Indeed, problem-solving researchers have investigated these types of interactions in the area of comprehension of science textbooks (Alexander and Kulikowich, 1991, 1994).

Lower-Order Problems May Not Require Content Knowledge, While Higher-Order Problems Promote Strong Content Usage

Because of the pervasive use among biology educators of Bloom's taxonomy to write and evaluate multiple-choice assessments, we decided it was valuable to examine the relationship between domain-general and domain-specific procedures and lower-order versus higher-order problems.

For both lower-order and higher-order problems, domain-specific procedures were used much more frequently than domain-general procedures (Table 5, A and B). This is comforting and unsurprising. We administered problems about ecosystems, evolution, and structure-function relationships, so

we expected and hoped students would use their knowledge of biology to solve these problems. However, two other results strike us as particularly important. First, domain-general procedures are highly prevalent (Table 5A, $n = 1108$ across all problems). The use of domain-general procedures is expected. There are certain procedures that are good practice in problem solving regardless of content, such as Analyzing Domain-General Visual Representations and Clarifying. However, students' extensive use of other domain-general/hybrid categories, namely Recognizing, is disturbing. Here we see students doing what all biology educators who use multiple-choice assessment fear, scanning the options for one that looks right based on limited knowledge. It is even more concerning that students' use of Recognizing is nearly four times more prevalent in lower-order problems than higher-order problems and that overall domain-general procedures are more prevalent in lower-order problems (Table 5A). As researchers have discovered, lower-order problems, not higher-order problems, are the type most often found in college biology courses (Momsen *et al.*, 2010). That means biology instructors' overreliance on lower-order assessment is likely contributing to students' overreliance on procedures that do not require biology content knowledge.

Second, it is striking that domain-specific procedures are more prevalent among higher-order problems than lower-order problems. These data suggest that higher-order problems promote strong content usage by students. As others have argued, higher-order problems should be used in class and on exams more frequently (Crowe *et al.*, 2008; Momsen *et al.*, 2010).

Using Domain-Specific Procedures May Improve Student Performance

Although it is interesting in and of itself to learn the procedures used by students during multiple-choice assessment, the description of these categories of procedures begs the question: does the type of procedure used by students make any difference in their ability to choose a correct answer? As explained in the *Introduction*, the strongest problem-solving approaches stem from a relatively complete and well-organized knowledge base within a domain (Chase and Simon, 1973; Chi *et al.*, 1981; Pressley *et al.*, 1987; Alexander and Judy, 1998). Thus, we hypothesized that use of domain-specific procedures would be associated with solving problems correctly, but use of domain-general procedures would not. Indeed, our data support this hypothesis. While limited use of domain-general procedures was associated with improved probability of success in solving multiple-choice problems, students who practiced extensive domain-specific procedures almost guaranteed themselves success in multiple-choice problem solving. In addition, as students used more domain-specific procedures, there was a weak but positive increase in the course performance, while use of domain-general procedures showed no correlation to performance. These data reiterate the conclusions of prior research that successful problem solvers connect information provided within the problem to their relatively strong domain-specific knowledge (Smith and Good, 1984; Pressley *et al.*, 1987). In contrast, unsuccessful problem solvers heavily depend on relatively weak domain-specific knowledge (Smith and Good, 1984; Smith, 1988). General problem-solving procedures can be used to make some progress in reaching a solution to domain-specific problems, but a problem solver can get only so

far with this type of thinking. In solving domain-specific problems, at some point, the solver has to understand the particulars of a domain to reach a legitimate solution (reviewed in Pressley *et al.*, 1987; Bassok and Novick, 2012). Likewise, problem solvers who misunderstand key conceptual pieces or cannot identify the deep, salient features of a problem will generate inadequate, incomplete, or faulty solutions (Chi *et al.*, 1981; Nehm and Ridgway, 2011).

Our findings strengthen the conclusions of previous work in two important ways. First, we studied problems from a wider range of biology topics. Second, we studied a larger population of students, which allowed us to use both qualitative and quantitative methods.

Limitations of This Research

Think-aloud protocols typically take place in an interview setting in which students verbally articulate their thought processes while solving a problem. When students are silent, the interviewer is there to prompt them to continue thinking aloud. We modified this protocol and taught students how to write out their procedures. However, one limitation of this study and all think-aloud studies is that it is not possible to analyze what students may have been thinking but did not state. Despite this limitation, we were able to identify a range of problem-solving procedures and errors that inform teaching and learning.

Implications for Teaching and Learning

There is general consensus among biology faculty that students need to develop problem-solving skills (NRC, 2003; AAAS, 2011). However, problem solving is not intuitive to students, and these skills typically are not explicitly taught in the classroom (Nehm, 2010; Hoskinson *et al.*, 2013). One reason for this misalignment between faculty values and their teaching practice is that biology problem-solving procedures have not been clearly defined. Our research presents a categorization of problem-solving procedures that faculty can use in their teaching. Instructors can use these well-defined problem-solving procedures to help students manage their knowledge of biology; students can be taught when and how to apply knowledge and how to restructure it. This gives students the tools to become more independent problem solvers (Nehm, 2010).

We envision at least three ways that faculty can encourage students to become independent problem solvers. First, faculty can model the use of problem-solving procedures described in this paper and have students write out their procedures, which makes them explicit to both the students and instructor. Second, models should focus on domain-specific procedures, because these steps improve performance. Explicit modeling of domain-specific procedures would be eye-opening for students, who tend to think that studying for recognition is sufficient, particularly for multiple-choice assessment. However, our data and those of other researchers (Stanger-Hall, 2012) suggest that studying for and working through problems using strong domain-specific knowledge can improve performance, even on multiple-choice tests. Third, faculty should shift from the current predominant use of lower-order problems (Momsen *et al.*, 2010) toward the use of more higher-order problems. Our data show that lower-order problems prompt for domain-general problem solving, while higher-order problems prompt for domain-specific problem solving.

We took what we learned from the investigation reported here and applied it to develop an online tutorial called SOLVEIT for undergraduate biology students (Kim *et al.*, 2015). In SOLVEIT, students are presented with problems similar to the stickleback problem shown in Figure 1. The problems focus on species concepts and ecological relationships. In brief, SOLVEIT asks students to provide an initial solution to each problem, and then it guides students through the problem in a step-by-step manner that encourages them to practice several of the problem-solving procedures reported here, such as Recalling, Checking, Analyzing Visual Representations, and Correcting. In the final stages of SOLVEIT, students are asked to revise their initial solutions and to reflect on an expert's solution as well as their own problem-solving process (Kim *et al.*, 2015). Our findings of improved student learning with SOLVEIT (Kim *et al.*, 2015) are consistent with the research of others that shows scaffolding can improve student problem solving (Lin and Lehman, 1999; Belland, 2010; Singh and Haileselassie, 2010). Thus, research to uncover the difficulties of students during problem solving can be directly applied to improve student learning.

ACKNOWLEDGMENTS

We thank the students who participated in this study and the biology faculty who served as experts by providing Bloom's rankings for each problem. We also thank the Biology Education Research Group at UGA, who improved the quality of this work with critical feedback on the manuscript. Finally, we thank the reviewers, whose feedback greatly improved the manuscript. Resources for this research were provided by UGA and the UGA Office of STEM Education.

REFERENCES

- Alexander PA, Judy JE (1988). The interaction of domain-specific and strategic knowledge and academic performance. *Rev Educ Res* 58, 375–404.
- Alexander PA, Kulikowich JM (1991). Domain-specific and strategic knowledge as predictors of expository text comprehension. *J Reading Behav* 23, 165–190.
- Alexander PA, Kulikowich JM (1994). Learning from physics text: a synthesis of recent research. *J Res Sci Teach* 31, 895–911.
- American Association for the Advancement of Science (2011). *Vision and Change in Undergraduate Biology Education: A Call to Action*, Washington, DC.
- Anderson C, Sheldon TH, Dubay J (1990). The effects of instruction on college non-majors' conceptions of respiration and photosynthesis. *J Res Sci Teach* 27, 761–776.
- Anderson LW, Krathwohl DR (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*, Boston, MA: Allyn & Bacon.
- Anderson TR, Schönborn KJ, du Plessis L, Gupthar AS, Hull TL (2012). Identifying and developing students' ability to reason with concepts and representations in biology. In: *Multiple Representations in Biological Education*, vol. 7, ed. DF Treagust and C Tsui, Dordrecht, Netherlands: Springer, 19–38.
- Bassok M, Novick LR (2012). Problem solving. In: *Oxford Handbook of Thinking and Reasoning*, ed. KJ Holyoak and RG Morrison, New York: Oxford University Press, 413–432.
- Belland BR (2010). Portraits of middle school students constructing evidence-based arguments during problem-based learning: the impact of computer-based scaffolds. *Educ Technol Res Dev* 58, 285–309.
- Bissell AN, Lemons PP (2006). A new method for assessing critical thinking in the classroom. *BioScience* 56, 66–72.
- Bloom BS (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals*, New York: McKay.
- Brown AL (1978). Knowing when, where, and how to remember: a problem of metacognition. In: *Advances in Instructional Psychology*, vol. 1, ed. R Glaser, Hillsdale, NJ: Erlbaum, 77–165.
- Brownell SE, Wenderoth MP, Theobald R, Okoroafor N, Koval M, Freeman S, Walcher-Chevillet CL, Crowe AJ (2014). How students think about experimental design: novel conceptions revealed by in-class activities. *BioScience* 64, 125–137.
- Bunce DM, Gabel DL, Samuel JV (1991). Enhancing chemistry problem-solving achievement using problem categorization. *J Res Sci Teach* 28, 505–521.
- Cartrette DP, Bodner GM (2010). Non-mathematical problem solving in organic chemistry. *J Res Sci Teach* 47, 643–660.
- Chase WG, Simon HA (1973). The mind's eye in chess. In: *Visual Information Processing*, ed. WG Chase, New York: Academic, 115–181.
- Chi MTH, Feltovich PJ, Glaser R (1981). Categorization and representation of physics problems by experts and novices. *Cogn Sci* 5, 121–152.
- Chi MTH, Glaser R (1985). Problem-solving ability. In: *Human Abilities: An Information-Processing Approach*, ed. RJ Sternberg, New York: Freeman.
- Crowe A, Dirks C, Wenderoth MP (2008). Biology in Bloom: implementing Bloom's taxonomy to enhance student learning in biology. *CBE Life Sci Educ* 7, 368–381.
- Dawson MRW (1998). *Understanding Cognitive Science*, 1st ed., Malden, MA: Wiley-Blackwell.
- Dirks C, Leroy C, Wenderoth MP (2013). Science Process and Reasoning Skills Test (SPARST): development and early diagnostic results. Presented at the Society for the Advancement of Biology Education Research (SABER) annual meeting, held 11–14 July 2013, in Minneapolis, MN.
- Duncker K, Lees LS (1945). On problem-solving. *Psychol Monogr* 58, i–113.
- Ericsson KA, Simon HA (1984). *Protocol Analysis: Verbal Reports as Data*, rev. ed., Cambridge, MA: MIT Press.
- Fitzmaurice GM, Laird NM, Ware JH (2011). *Applied Longitudinal Analysis*, 2nd ed., Hoboken, NJ: Wiley.
- Gormally C, Brickman P, Lutz M (2012). Developing a test of scientific literacy skills (TOSLS): measuring undergraduates' evaluations of scientific information and arguments. *CBE Life Sci Educ* 11, 364–377.
- Halpern DE (1997). *Critical Thinking across the Curriculum: A Brief Edition of Thought and Knowledge*, Mahwah, NJ: Erlbaum.
- 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.
- Hoskinson A-M, Caballero MD, Knight JK (2013). How can we improve problem solving in undergraduate biology? Applying lessons from 30 years of physics education research. *CBE Life Sci Educ* 12, 153–161.
- Jacobs JE, Paris SG (1987). Children's metacognition about reading—issues in definition, measurement, and instruction. *Educ Psychol* 22, 255–278.
- Jonassen D (2012). Designing for problem solving. In: *Trends and Issues in Instructional Design and Technology*, 3rd ed., ed. RA Reiser and JV Dempsey, Boston, MA: Pearson Education, 64–74.
- Jonassen DH (2000). Toward a design theory of problem solving. *Educ Technol Res Dev* 48, 63–85.
- Keys CW (2000). Investigating the thinking processes of eighth grade writers during the composition of a scientific laboratory report. *J Res Sci Teach* 37, 676–690.
- Kim HS, Prevost L, Lemons PP (2015). Students' usability evaluation of a Web-based tutorial program for college biology problem solving. *J Comput Assist Learn* 31, 362–377.
- Kohl PB, Finkelstein ND (2008). Patterns of multiple representation use by experts and novices during physics problem solving. *Phys Rev Spec Top Phys Educ Res* 4, 010111.
- Lavoie DR (1993). The development, theory, and application of a cognitive-network model of prediction problem solving in biology. *J Res Sci Teach* 30, 767–785.
- Lin X, Lehman JD (1999). Supporting learning of variable control in a computer-based biology environment: effects of prompting college students to reflect on their own thinking. *J Res Sci Teach* 36, 837–858.
- Martinez ME (1998). What is problem solving? *Phi Delta Kappan* 79, 605–609.

- Momsen J, Offerdahl E, Kryjevskaja M, Montplaisir L, Anderson E, Grosz N (2013). Using assessments to investigate and compare the nature of learning in undergraduate science courses. *CBE Life Sci Educ* 12, 239–249.
- Momsen JL, Long TM, Wyse SA, Ebert-May D (2010). Just the facts? Introductory undergraduate biology courses focus on low-level cognitive skills. *CBE Life Sci Educ* 9, 435–440.
- National Research Council (2003). *BIO2010: Transforming Undergraduate Education for Future Research Biologists*, Washington, DC: National Academies Press.
- Nehm RH (2010). Understanding undergraduates' problem-solving processes. *J Microbiol Biol Educ* 11, 119–122.
- Nehm RH, Reilly L (2007). Biology majors' knowledge and misconceptions of natural selection. *BioScience* 57, 263–272.
- Nehm RH, Ridgway J (2011). What do experts and novices "see" in evolutionary problems? *Evol Educ Outreach* 4, 666–679.
- Newell A, Shaw JC, Simon HA (1958). Elements of a theory of human problem solving. *Psychol Rev* 65, 151–166.
- Newell A, Simon HA (1972). *Human Problem Solving*, Upper Saddle River, NJ: Prentice-Hall.
- Patton MQ (1990). *Qualitative Evaluation and Research Methods*, Thousand Oaks, CA: Sage.
- Polya G (1957). *How to Solve It*, Garden City, NY: Doubleday.
- Pressley M, Borkowski JG, Schneider W (1987). Cognitive strategies: good strategy users coordinate metacognition and knowledge. *Ann Child Dev* 4, 89–129.
- Pressley M, Goodchild F, Fleet J, Zajchowski R, Evans ED (1989). The challenges of classroom strategy instruction. *Elem Sch J* 89, 301–342.
- Runco MA, Chand I (1995). Cognition and creativity. *Educ Psychol Rev* 7, 243–267.
- Schraw G, Moshman D (1995). Metacognitive theories. *Educ Psychol Rev* 7, 351–371.
- Singer SR, Nielsen NR, Schweingburger HA, Committee on the Status, Contributions, and Future Directions of Discipline-Based Education Research (2012). *Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering*, Washington, DC: National Academies Press.
- Singh C, Haileselassie D (2010). Developing problem-solving skills of students taking introductory physics via Web-based tutorials. *J Coll Sci Teach* 39, 42–49.
- Smith JI, Combs ED, Nagami PH, Alto VM, Goh HG, Gourdet MA, Hough CM, Nickell AE, Peer AG, Coley JD, et al. (2013). Development of the biology card sorting task to measure conceptual expertise in biology. *CBE Life Sci Educ* 12, 628–644.
- Smith MU (1988). Successful and unsuccessful problem solving in classical genetic pedigrees. *J Res Sci Teach* 25, 411–433.
- Smith MU (1992). Expertise and the organization of knowledge: unexpected differences among genetic counselors, faculty, and students on problem categorization tasks. *J Res Sci Teach* 29, 179–205.
- Smith MU, Good R (1984). Problem solving and classical genetics: successful versus unsuccessful performance. *J Res Sci Teach* 21, 895–912.
- Stanger-Hall KF (2012). Multiple-choice exams: an obstacle for higher-level thinking in introductory science classes. *CBE Life Sci Educ* 11, 294–306.
- White RT (1998). Decisions and problems in research on metacognition. In: *International Handbook of Science Education*, ed. B Fraser and KG Tobin, Dordrecht, Netherlands: Springer, 1207–1213.
- Zheng AY, Lawhorn JK, Lumley T, Freeman S (2008). Application of Bloom's taxonomy debunks the "MCAT myth." *Science* 319, 414–415.
- Zoller U (1993). Are lecture and learning compatible? Maybe for LOCS: unlikely for HOCS. *J Chem Educ* 70, 195–197.