Student Performance along Axes of Scenario Novelty and Complexity in Introductory Biology: Lessons from a Unique Factorial Approach to Assessment

Kirsten K. Deane-Coe,[†‡§*] Mark A. Sarvary,[‡] and Thomas G. Owens[†]
[†]School of Integrative Plant Sciences and [‡]Investigative Biology Laboratories, Department of Neurobiology and Behavior, Cornell University, Ithaca, NY 14853

ABSTRACT

In an undergraduate introductory biology laboratory course, we used a summative assessment to directly test the learning objective that students will be able to apply course material to increasingly novel and complex situations. Using a factorial framework, we developed multiple true–false questions to fall along axes of novelty and complexity, which resulted in four categories of questions: familiar content and low complexity (category A); novel content and low complexity (category B); familiar content and high complexity (category C); and novel content and high complexity (category D). On average, students scored more than 70% on all questions, indicating that the course largely met this learning objective. However, students scored highest on questions in category A, likely because they were most similar to course content, and lowest on questions in categories C and D. While we anticipated students would score equally on questions for which either novelty or complexity was altered (but not both), we observed that student scores in category C were lower than in category B. Furthermore, students performed equally poorly on all questions for which complexity was higher (categories C and D), even those containing familiar content, suggesting that application of course material to increasingly complex situations is particularly challenging to students.

INTRODUCTION

Postsecondary introductory biology courses commonly seek to teach application of analytical techniques in addition to requisite content knowledge. Typically, students are also expected to be able to apply knowledge gained in introductory courses in upper-division electives or in laboratory settings. Introductory biology courses thus carry the responsibility of not only teaching students course material, but also emphasizing the acquisition of knowledge in a way that allows students to directly apply that knowledge in research contexts (Mintzes and Wandersee, 1997; Michael, 2001), thus preparing them for the diversity of next steps in their academic careers (American Association for the Advancement of Science, 2011). For students, successful application of skills in different contexts requires the application of analytical techniques across disciplinary boundaries, in novel biological systems, and in increasingly complex situations.

Application of course content in diverse contexts, however, may pose challenges for students. There is substantial evidence that learners experience difficulty solving problems that involve deviation from procedures on which they have been trained (Reed *et al*., 1985; Ross, 1987, 1989; Novick and Holyoak, 1991; Catrambone, 1994, 1995, 1996). Novel problem solving requires departure from routine conceptual tasks and significant cognitive searching (Newell, 1980; Anderson, 1993), which imposes cognitive demands on students. Successful transfer of learning to novel or more complex problems depends on the degree and type of existing knowledge and the instructional

Joel K. Abraham, *Monitoring Editor*
Submitted June 20, 2016; Revised November 4, 2016; Accepted November 14, 2016
CBE Life Sci Educ March 1, 2017 16:ar3
DOI:10.1187/cbe.16-06-0195

[§]Present address: Department of Biology, St. Mary's College of MD, St. Mary's City, MD 20686.
*Address correspondence to: Kirsten K. Deane-Coe (kkdeanecoe@smcm.edu).

design context in which the problems are presented (Reed and Saavedra, 1986; Barnett and Ceci, 2002). Further, novel and complex material may impose a high cognitive load (total amount of mental effort being used in working memory) for students when they encounter such situations (Sweller, 1988; Sweller and Chandler, 1991, 1994; Van Merriernboer and Sweller, 2005), which may influence their ability to successfully complete problems.

If the ability to apply course content in diverse ways is a central goal in introductory biology, and likely an implicit learning objective in many introductory biology courses, two related questions that arise are as follows: 1) How can we determine whether the ability to apply course content is achieved in introductory biology courses? 2) In what specific ways is application of course content most challenging for students? The answers to these questions can help guide course development in biology to meet the needs of a student body with an increasingly diverse set of academic and career paths.

Assessments can provide a means of investigating the questions posed above. Through alignment of course activities, assessments can be created to test student performance on achieving course learning objectives (Boud and Falchikov, 2006; Handelsman *et al*., 2007). One successful assessment technique that can directly address the application of course content in diverse contexts is interleaved questioning, in which students are tasked to apply a similar concept to a variety of different situations (Rohrer, 2012). A second approach is a scaffolded approach, in which question complexity is increased as students gain confidence with basic concepts. This technique was originally developed using mentor and peer interactions to provide feedback and support during more complex learning tasks (King *et al*., 1997), but it has recently been explored in software-based learning environments (Reiser, 2004) and can be adapted to summative assessments. In both of these approaches, students are required to apply learned course content to different situations, either using lateral, interleaved designs or incremental, scaffolded designs.

In this study, we focused on an introductory biology laboratory course and assessed the ability of students to apply course content in diverse ways. This introductory course was designed to 1) expose students to realistic scientific questions where they design hypothesis-based experiments, choose appropriate statistical test(s), analyze data and interpret results; 2) fill students' scientific "toolbox" by demonstrating mastery of modern lab techniques and scientific methods that will be applied across biological systems and scales; 3) teach students how to think through a scientific process with their research group while acquiring conceptual knowledge and understanding the benefits and challenges of collaborative work; 4) teach students how to find relevant scientific information using appropriate library tools and how to communicate effectively using both written and oral formats; and 5) guide students as they apply course material to increasingly novel and complex situations. In this study, we examine the success of the course in meeting goal 5, which was also a student learning objective (students will be able to apply course material to novel and complex situations), using a factorial approach containing interleaved and scaffolded questions within a summative assessment. Specifically, we designed questions that increased novelty, complexity, or both (compared with material taught in the course) and compared scores to questions that did neither (Figure 1). We defined novelty as a conceptual scenario not directly encountered by students in the course (e.g., a different organism or biological system), and we defined complexity as an increase in the number of variables presented in a problem (e.g., two response variables presented compared with one).

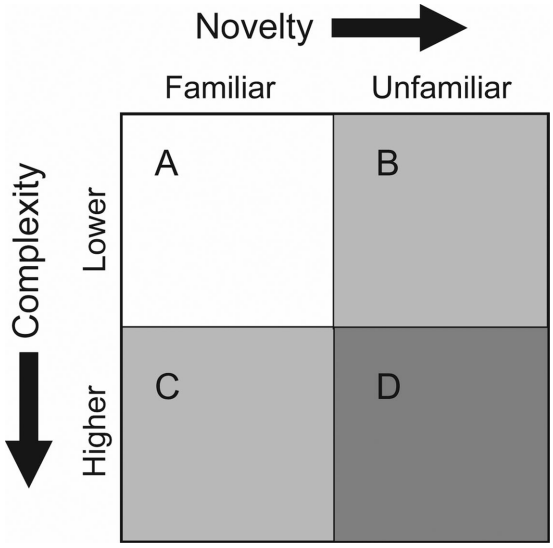

FIGURE 1. Factorial framework used for generating and categorizing questions in four categories (A–D) along axes of novelty and complexity.

We predicted that student performance would be highest for questions with familiar content and complexity (category A), and lowest for questions with higher complexity *and* novel content (category D). We also predicted that student performance would be reduced equally (and to an intermediate degree) in contexts in which students were faced with either novel (category B) *or* more complex (category C) questions.

METHODS

Course Description

Questions were administered to students in a semester-long introductory biology laboratory course for biology majors, called Investigative Biology (BioG1500), at Cornell University. BioG1500 includes one 55-minute lecture and one 3-hour laboratory period per week. The course typically enrolls 350–400 students per semester (24 lab sections of 15–18 students each) and is taught in both Spring and Fall semesters. The course was cotaught by the same two instructors for the three consecutive semesters of data used in this study. The course is run in a modular format focusing on content areas of 1) selection and antibiotic resistance in bacteria; 2) freshwater ecology and nutrient limitation in phytoplankton; and 3) human microsatellite DNA. Active-learning techniques such as case studies, clicker questions, problem sequences, and think–pair–share exercises are used throughout lectures. The lab portion of the course focuses on hands-on experiments, lab skills, statistical reasoning, data interpretation, and exposure to the process of science, from hypothesis development to paper/poster production. (See http://investigativebiology.cornell.edu for additional course information.)

Assessment Development and Validation

To test students' ability to apply conceptual and analytical knowledge taught in the course to 1) novel and 2) increasingly complex situations, we created questions organized along these two axes that followed a factorial design (Figure 1). Questions were developed to fit into four categories: questions that contained familiar content and lower complexity (A), questions that contained unfamiliar content and lower complexity (B), questions that contained familiar content and higher complexity (C), and questions that contained unfamiliar content and higher complexity (D). Please refer to the Supplemental Material for examples of all questions administered. To develop questions along the novelty axis, we used a novel conceptual version of a familiar type of figure or data table in the questions. To develop questions along the complexity axis, we increased the number of variables in a figure or data table by one in the questions (e.g., three variables in a C question compared with two variables in an A question) (see Van der Meij and de Jong, 2006).

All questions were in the multiple true–false format, which required students to evaluate individual statements and provided an additional layer of assessment compared with standard, one-answer multiple-choice questions. Questions were organized such that, for each category (A–D), students encountered a figure or a table containing data to interpret, a question stem, and eight true–false statements. Question stems and associated true–false statements were changed to meet the criteria for categories A–D. We created two complete sets of questions (two sets of eight multiple true–false statements per category) for two reasons: first, to reduce the chance that a single group of questions or particular premise would influence performance in a category, and second, to ensure mixing of questions among the 24 laboratory sections of the course. All students received two question stems from one set (categories A and B) and two question stems from the other set (categories C and D), so learning occurring within question sets did not play a role in student responses. To ensure that all students were assessed on similar course content, we were careful to include types of true–false statements across the question sets that addressed a similar range of concepts, such as data interpretation, understanding of statistical analyses presented in R outputs, and inference of evolutionary processes.

For validation of the four categories in which questions were developed to fit, copies of the questions were given to current graduate and undergraduate teaching assistants (TAs) for the course (n = 20), who were charged to identify the category (A–D) in which each question belonged. The TAs were not given any information about the question categories before this procedure, allowing us to objectively assess fit in each of the categories. We used current TAs for this exercise, because questions in categories B–D were developed in reference to questions in category A (containing familiar content and lower complexity); thus, their relative placement in the categories required conceptual knowledge of course material similar to that of students. TAs correctly placed questions in categories A, B, C, and D 74, 71, 62, and 53% of the time, respectively.

Administering Questions in the Course

Questions in all categories were provided to students as part of a laboratory practical at the end of each semester for three consecutive semesters. Students were able to practice the multiple true–false format on other questions not part of this study in the first lab practical of the course. Thus, students were familiar with the question format before the final practical. Paper copies of four groups of eight questions organized by category (A–D) were placed around the room at different stations, similar to the other questions on the practical exam. Students selected each of the four question groups in no particular order and at the time of their choosing during the practical exam. Students answered the questions individually at their desks by filling in their answers for all 32 questions on a Scantron (Eagan, MN) sheet. Questions were administered in three semesters for a total of 1139 participants. For each student, the number of questions scored correct out of eight for each category was calculated, and this number was used as a continuous response variable in subsequent analyses.

Data Analysis

A discrimination analysis was used to determine the relative strength of questions in their ability to discriminate between higher- and lower-performing students on the final practical exam. We used the following formula (see Wood, 1960; Ebel and Frisbie, 1986; Wiersma and Jurs, 1990; Ding and Biechner, 2009) to calculate the discrimination index (DI) value for each question:

$$\mathrm{DI} = (U_P - L_P) / N$$

Students were ranked based on their practical exam grades and were subcategorized into an upper group representing the highest 30% of scores and a lower group representing the lowest 30% of scores. For each question, U_p represents the number of students in the upper group who answered the question correctly, L_p represents the number of students in the lower group who answered the question correctly, and N represents the total number of students in the upper group. Values of DI for individual questions thus range from −1, indicating that all students in the lower group answered the question correctly, to +1, indicating that all students in the upper group answered the question correctly. In general, using this metric, the higher the DI value, the greater the ability of the question to discriminate among higher- and lower-performing students.

The resulting DI values were used in conjunction with ease index values (proportion of students answering a particular question correctly) to identify any questions that needed to be eliminated because students may have been getting them wrong for the wrong reasons. In particular, questions with DI values below zero and ease index values below 50 were heavily scrutinized. This combination indicated 50% or fewer students answered the question correctly, and among those, more students in the lower-performing group answered the question correctly than did students in the upper-performing group. We conducted this quality-control analysis in the semester before the three presented here to determine whether all questions met the desired criteria. Several questions were subsequently altered for clarity or content before testing in the three semesters used in this study, and all questions for which data are presented met the desired quality criteria.

We used a mixed-effects model to examine the ability of students to answer questions in the four categories (A–D) that were organized along the two axes: novelty and complexity.

The main effects in the model included novelty, complexity, question set (1 or 2), and semester. The interaction between novelty and complexity was also included in the model, and student nested in section was included as a random effect. All analyses were performed using the open-source platform R (R Core Team, 2015) and used the statistical cutoff of $\alpha = 0.05$.

RESULTS

DI values did not differ across categories (A–D), question sets (1, 2), or semesters (Spring 2014, S14; Fall 2014, F14; and Spring 2015, S15). When data from all semesters were combined, mean (± SD) DI values for question categories A, B, C and D were 0.13 (± 0.11), 0.16 (± 0.10), 0.15 (± 0.09), and 0.12 (± 0.07), respectively (Figure 2).

On average, students scored above 70% on all questions, but scores differed across categories. Students correctly answered 78% of questions in category A, scores were reduced by 7.5% in category B, and by 8% in categories C and D (Figure 3). Across the three semesters analyzed, there was a positive relationship between scores on multiple true–false questions administered in the exam (32 points total) and scores on the entire exam (Supplemental Figure S1).

The mixed-effects model revealed that complexity and the interaction between novelty and complexity (novelty × complexity) were both significant predictors of student scores on questions (Table 1). When the interaction between novelty and complexity was analyzed in more detail, it was apparent that complexity was a greater determinant of student performance when students encountered familiar (nonnovel) questions versus unfamiliar (novel) questions. Scores on low-complexity questions were 8% higher than high-complexity questions when the subject matter was familiar, but only 3% higher when the subject matter was novel (Figure 4A). In contrast, scores on questions with familiar subject matter were 7.5% higher compared with novel subject matter (Figure 4B). Therefore, novelty did not influence student performance when complexity was high, but did when complexity was low.

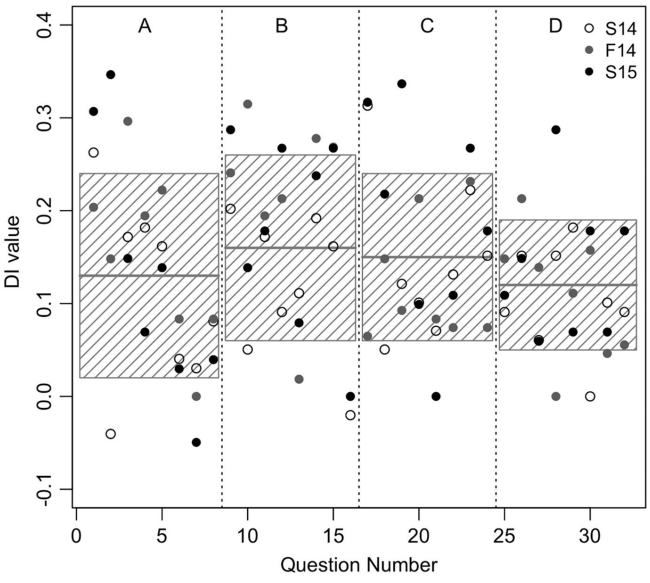

FIGURE 2. Discrimination index (DI) values for questions in categories A–D administered in Spring 2014 (S14, open symbols), Fall 2014 (F14, gray symbols), and Spring 2015 (S15, black symbols). DI values did not differ significantly across categories or semesters ($p > 0.05$). Gray horizontal lines and shaded boxes represent the mean ± SD for each question category.

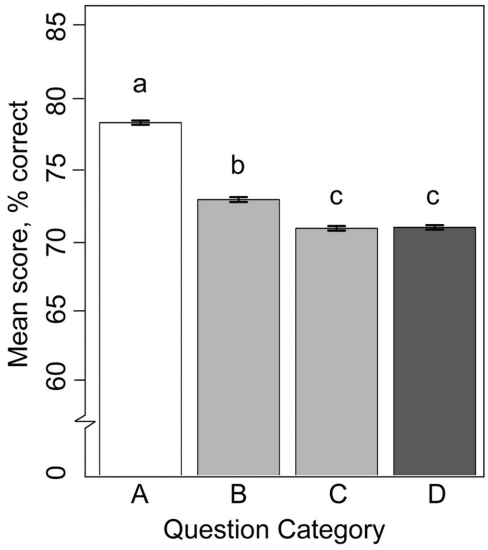

FIGURE 3. Mean (± SE) student scores (% correct) for questions in categories A–D over three semesters ($n = 1139$). Data were pooled from the three semesters to illustrate trends because semester was not a significant main effect in the mixed-effects model used. Different lowercase letter represent significant differences between categories ($p < 0.05$).

DISCUSSION

In this study, we used a summative assessment to examine students' ability to apply material learned in an introductory biology course to increasingly novel and more complex situations. In support of our first prediction, students scored highest on questions with familiar content and complexity (category A). As predicted, students scored lowest on questions with higher complexity and novel content (category D), but they also scored equivalently low on questions with higher complexity and familiar content (category C). Our second prediction, that students would be challenged equally by questions of higher complexity and novelty (categories B and C), was not supported. Instead, we found that progressing to a higher complexity

TABLE 1. Mixed-effects model to explain the drivers of student scores on multiple true–false questions[a]

Fixed effect	Parameter value	SE	*t*	*p*[b]
Novelty	0.001	0.0067	0.137	0.891
Complexity	0.079	0.0067	11.74	**<0.001**
Semester	0.001	0.0065	0.570	0.305
Question set	0.004	0.0089	0.411	0.681
Novelty × complexity	−0.058	0.0095	−6.147	**<0.001**

[a]Novelty, complexity, their interaction (novelty × complexity), semester, and question set were included as fixed effects, and student nested within lab section was included as a random effect.
[b]Values in bold indicate significant effects ($p < 0.05$).

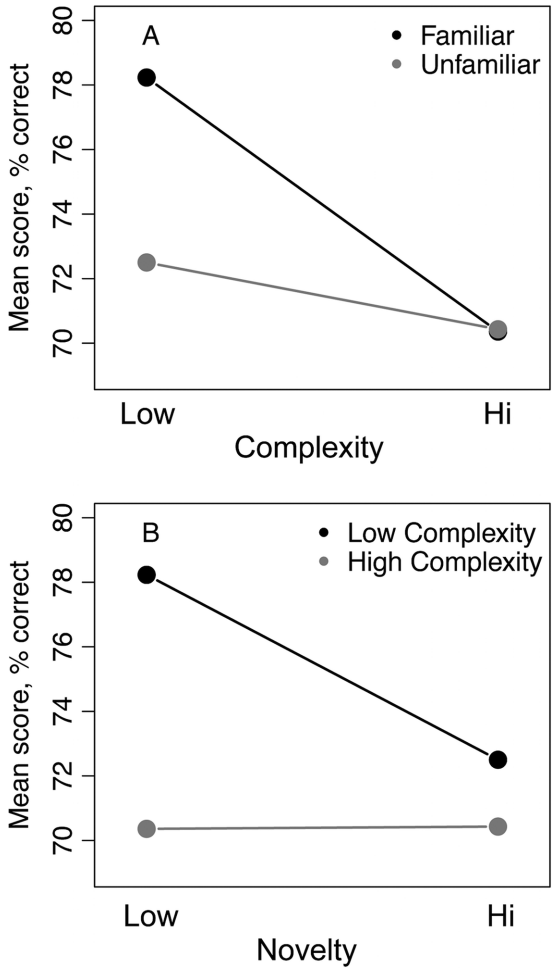

FIGURE 4. Interaction plots for the effect of complexity on scores for questions that differed in novelty (A), and the effect of novelty on scores for questions that differed in complexity (B). Points indicate the mean values ($n = 1139$) in each of the categories, and changes in slope of the connecting lines indicate the strength of interaction between the factors.

reduced student performance to a greater extent than moving to novel subject matter. Student scores on questions were highest when questions contained familiar and low-complexity situations (A) and lowest when questions contained high-complexity situations (C and D), regardless of the novelty of the subject matter.

Variation in student performance based on question category can be partially explained by the familiarity level of exam questions, both in complexity and novelty. Over the course of the semester, via formative assessments and in-class problems, students were exposed to situations that challenged them to apply knowledge laterally in novel contexts and in situations that increased in complexity, but it is possible that this exposure, particularly with respect to complexity, was insufficient to develop proficiency in these skills. With respect to exposure to novelty, students used the scientific method to develop hypotheses and design experiments using different biological topics and organisms in all three course modules. Because the scientific process was a central theme in the course, it is likely that repeated exposure to novel situations in this manner resulted in students struggling to a lesser degree on exam questions in category B (scores reduced by 7.5%) compared with those in categories C and D (scores reduced by 8%). With respect to complexity, students were challenged to apply increasingly sophisticated statistical techniques to data sets and to develop proficiency in writing and interpreting increasingly complicated code in R (R Core Team, 2015). However, it is possible that the emphasis on developing proficiency with increasingly complex situations was too limited, both in scope and time allocation, to develop proficiencies in application of these skills.

While this study was the first to directly test the effects of novelty and complexity in a biology course, reduced student performance on novel and more complex tasks has been reported in linguistics and physical sciences. Novelty has been shown to reduce performance at second language comprehension (Schmidt-Rinehart, 1994), but only in the case of listening (not reading). However, exposure to novel conceptual tasks within a similar level of complexity in an interleaved design has been shown to increase students' ability to apply knowledge in conceptually related contexts (Rohrer, 2012). Increased complexity has been shown to influence student performance in a simulation-based physics learning environment, but its effect depended on how concepts were represented in simulations (Van der Meij and de Jong, 2006). These contrasting results demonstrate the need for factorial designs such as the one applied in this study that address novel and complex situations compared with a course-based control group (category A in our study).

One surprising outcome of this study was the strong effect of question complexity on student performance. Students did not perform equally on questions for which either novelty or complexity was altered, and scores on assessments were influenced by the complexity level of the question rather than the novelty of the subject matter. At the high-complexity level, scores were equally reduced compared with scores on questions in category A, even when subject matter was familiar. We also discovered that, while complexity of subject matter influenced student scores regardless of novelty, novelty of subject matter only mattered for lower-complexity questions. This suggests students were challenged by increases in complexity across the board but were only differentially challenged by novel situations when questions were simpler. In contrast to these results, in a physics course exam, Van der Meij and de Jong (2006) discovered differences in student scores between treatments (different types of representations of questions) and student perception of differences in difficulty only when question complexity was high. One difference between this study and the present study was that four scaffolded levels were used by Van der Meij and de Jong (2006) compared with two levels used here (e.g., A versus C questions). These contrasting results may therefore relate to the number of scaffolded levels used in the questioning scheme or the familiarity of scaffolded assessments to students in a given course.

Challenges associated with increasingly complex tasks in assessments are likely related to their degree of complexity compared with the degree of complexity in material encountered by students during a course (Catrambone, 1998). Cortright *et al*. (2005) showed that the relationship between performance on questions and task complexity is negative but

nonlinear, and a sharp decline in performance occurs after a threshold of complexity level. Also, the manner in which complex examples are presented and practiced by students may influence student performance on summative assessments. For example, breaking down complex problems into modular units before solving problems has been shown to facilitate successful problem solving for students compared with less structured approaches (Gerjets *et al*., 2004). Finally, student perception of the complexity of a task can also influence performance on a question. When questions are perceived to be more complex, they are more frequently answered incorrectly than those perceived to be simpler (Lee, 2004). We did not directly assess student perception of difficulty in the questions administered in this study, though combining quantitative and qualitative feedback in future work would give insight into how perceived difficulty influences student scores, as well as what components of questions led students to perceive them as more or less difficult. Further, considering student perception of difficulty level of questions on a numerical scale would provide an additional variable to potentially explain levels of correctness within and between question categories.

One possible explanation for the negative effect of complexity on student scores is that students may require additional learning strategies before they are prepared to accomplish higher-complexity cognitive tasks. In a psychology course, De Koning *et al*. (2007) found that visual cuing was essential in meeting learning outcomes using animations of high complexity. This suggests that forms of priming, particularly using the same sensory format, may enhance student ability to apply concepts of familiar complexity to higher-complexity tasks. In this study, we investigated two different types of cognitive tasks: those that were incremental (increasing complexity) and those that were lateral (increasing novelty). It is possible that each of these tasks required different learning strategies and course work succeeded more at developing the lateral framework compared with the incremental one.

Several considerations must be taken into account regarding the level of inference that can be drawn from the current study. First, while course material did include novel (more prominently) and more complex (less prominently) situations for students to grapple with, the majority of assessment techniques experienced by students throughout the semester related directly to course material and contained familiar content and complexity levels. For this reason, it is likely that the familiar and lower-complexity questions in category A were most similar to what they had experienced during the course compared with questions in the other categories. The relatively high performance on category A questions was probably partly due to cognitive mechanisms related to repetition and recall (Roediger and Karpicke, 2006).

Second, the question validation method used, including current course TAs, resulted in variable placement of questions into categories. While 74% of TAs correctly placed questions into category A, only 53% correctly placed the questions into category D. The TAs came from different disciplinary backgrounds (e.g., ecology, evolutionary biology, natural resources, entomology), and both undergraduate and graduate TAs were used for this assessment. The variability in general ability to assess questions based on these differences among TAs could have contributed to the apparent challenges in category placement we observed, particularly in the questions that contained higher-complexity and novel situations. We used an anonymous, Scantron-based question validation procedure and did not ask TAs to self-identify as undergraduate or graduate, though this would be an important component of follow-up studies. Nonetheless, while it would be optimal for agreement percentages to be higher, this study was the first of its kind, and we feel these presented data represent a logical baseline to which other courses or assessments can be compared.

In conclusion, this study illustrates how summative assessments are useful tools for direct evaluation of course learning objectives and for evaluating student performance with novel and more complex problems. Using a 2×2 factorial question design in which combinations of novelty and complexity were represented, we discovered that the interaction of these two variables explained variability in student scores on a final exam. Specifically, we found that progressing to questions of a higher complexity reduced student performance to a greater extent than moving to questions of novel subject matter. Returning to the research questions posed earlier in the context of the current course (How can we determine whether the goal of application of course content in diverse ways is met in introductory biology courses?, Which elements of application of course content are most challenging for students?), our data indicate that there was better student performance in demonstrating application of course content to novel situations than skills that allow students to progress to incrementally more complex situations. We show here that increasing the complexity of questions poses a significant challenge to students in introductory biology. Students may require additional cues or learning strategies to make incremental cognitive steps with more complex questions, while novel situations may pose a less challenging lateral step. We propose that scaffolded questions that gradually increase in complexity should be integrated into activities in introductory biology courses to enable students to apply learned material to increasingly complex situations they are likely to encounter in their academic careers.

ACCESSING MATERIALS

All questions used in the study are available in PDF format in the Supplemental Material.

ACKNOWLEDGMENTS

We thank Elise West for valuable feedback and discussions on data gathered in this study, BioG1500 graduate and undergraduate TAs (2014–2015) for assistance with question validation, BioG1500 staff members K. C. Bennett and Martha Lyon for logistical help, the Cornell Office of Undergraduate Biology for support of research, and the valuable and extensive feedback of two anonymous reviewers of the manuscript.

REFERENCES

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

Anderson JR (1993). Problem solving and learning. Am Psychol 48, 35.

Barnett SM, Ceci SJ (2002). When and where do we apply what we learn? A taxonomy for far transfer. Psychol Bull 128, 612.

Boud D, Falchikov N (2006). Aligning assessment with long-term learning. Assess Eval High Educ 31, 399–413.

Catrambone R (1994). Improving examples to improve transfer to novel problems. Mem Cogn 22, 606–615.

Catrambone R (1995). Aiding subgoal learning: effects on transfer. J Educ Psychol 87, 5–17.

Catrambone R (1996). Generalizing solution procedures learned from examples. J Exp Psychol Learn Mem Cogn 22, 1020–1031.

Catrambone R (1998). The subgoal learning model: creating better examples so that students can solve novel problems. J Exp Psychol Gen 127, 355.

Cortright RN, Collins HL, DiCarlo SE (2005). Peer instruction enhanced meaningful learning: ability to solve novel problems. Adv Physiol Educ 29, 107–111.

de Koning BB, Tabbers HK, Rikers RMJP, Paas F (2007). Attention cueing as a means to enhance learning from an animation. Appl Cogn Psychol 21, 731–746.

Ding L, Beichner R (2009). Approaches to data analysis of multiple-choice questions. Phys Rev Spec Top Phy Educ Res 5, 020103.

Ebel RL, Frisbie DA (1986). Essentials of Educational Measurement, Englewood Cliffs, NJ: Prentice-Hall.

Gerjets P, Scheiter K, Catrambone R (2004). Designing instructional examples to reduce intrinsic cognitive load: molar versus modular presentation of solution procedures. Instr Sci 32(1–2), 33–58.

Handelsman J, Miller S, Pfund C (2007). Scientific Teaching, New York: W. H. Freeman.

King A, Staffieri A, Adelgais A (1997). Mutual peer tutoring: effects of structuring tutorial interaction to scaffold higher level complex learning. J Educ Psychol 90, 134–152.

Lee Y (2004). Student perceptions of problems' structuredness, complexity, situatedness, and information richness and their effects on problem-solving performance. Doctoral Dissertation, Tallahassee: College of Education, Florida State University.

Michael J (2001). In pursuit of meaningful learning. Adv Physiol Educ 25, 145–158.

Mintzes JJ, Wandersee JH (1997). Reform and innovation in science teaching: a human constructivist view. In: Teaching Science for Understanding, ed. JJ Mintzes, JH Wandersee, and JD Novak, San Diego, CA: Academic, 29–58.

Newell A (1980). Reasoning, problem-solving, and decision processes: the problem space as a fundamental category. In: Attention and Performance VIII, ed. R Nickerson, Hillsdale, NJ: Erlbaum, 693–718.

Novick LR, Holyoak KJ (1991). Mathematical problem solving by analogy. J Exp Psychol Learn Mem Cogn 17, 398–415.

R Core Team (2015). R: A Language and Environment for Statistical Computing, Vienna, Austria: R Foundation for Statistical Computing. www.R-project.org (accessed May 2016).

Reed SK, Dempster A, Ettinger M (1985). Usefulness of analogous solutions for solving algebra word problems. J Exp Psychol Learn Mem Cogn 11, 106–125.

Reed SK, Saavedra NC (1986). A comparison of computation, discovery, and graph procedures for improving students' conception of average speed. Cogn Instr 3, 31–62.

Reiser BJ (2004). Scaffolding complex learning: the mechanisms of structuring and problematizing student work. J Learn Sci 13, 273–304.

Roediger HL, Karpicke JD (2006). The power of testing memory: basic research and implications for educational practice. Perspect Psychol Sci 1, 181–210.

Rohrer D (2012). Interleaving helps students distinguish among similar concepts. Educ Psychol Rev 24, 355–367.

Ross BH (1987). This is like that: the use of earlier problems and the separation of similarity effects. J Exp Psychol Learn Mem Cogn 13, 629–639.

Ross BH (1989). Distinguishing types of superficial similarities: different effects on the access and use of earlier problems. J Exp Psychol Learn Mem Cogn 15, 456–468.

Schmidt-Rinehart BC (1994). The effects of topic familiarity on second language listening comprehension. Mod Lang J 78, 179–189.

Sweller J (1988). Cognitive load during problem solving: effects on learning. Cogn Sci 12, 257–285.

Sweller J, Chandler P (1991). Evidence for cognitive load theory. Cogn Instr 8, 351–362.

Sweller J, Chandler P (1994). Why some material is difficult to learn. Cogn Instr 12, 185–233.

Van der Meij J, de Jong T (2006). Supporting students' learning with multiple representations in a dynamic simulation-based learning environment. Learn Instr 16, 199–212.

Van Merrienboer JJ, Sweller J (2005). Cognitive load theory and complex learning: recent developments and future directions. Educ Psychol Rev 17, 147–177.

Wiersma W, Jurs SG (1990). Educational Measurement and Testing, 2nd ed., Boston, MA: Allyn and Bacon.

Wood DA (1960). Test Construction: Development and Interpretation of Achievement Tests, Columbus, OH: Charles E. Merrill.