CBE—Life Sciences Education
Vol. 12, 47–58, Spring 2013

Article

Questions for Assessing Higher-Order Cognitive Skills: It's Not Just Bloom's

Paula P. Lemons* and J. Derrick Lemons†

*Department of Biochemistry and Molecular Biology and †Departments of Anthropology and Religion, University of Georgia, Athens, GA 30602

Submitted March 2, 2012; Revised September 24, 2012; Accepted September 25, 2012
Monitoring Editor: Robert L. DeHaan

We present an exploratory study of biologists' ideas about higher-order cognition questions. We documented the conversations of biologists who were writing and reviewing a set of higher-order cognition questions. Using a qualitative approach, we identified the themes of these conversations. Biologists in our study used Bloom's Taxonomy to logically analyze questions. However, biologists were also concerned with question difficulty, the length of time required for students to address questions, and students' experience with questions. Finally, some biologists demonstrated an assumption that questions should have one correct answer, not multiple reasonable solutions; this assumption undermined their comfort with some higher-order cognition questions. We generated a framework for further research that provides an interpretation of participants' ideas about higher-order questions and a model of the relationships among these ideas. Two hypotheses emerge from this framework. First, we propose that biologists look for ways to measure difficulty when writing higher-order questions. Second, we propose that biologists' assumptions about the role of questions in student learning strongly influence the types of higher-order questions they write.

INTRODUCTION

A number of national reports call for college science instructors to teach in a way that promotes the application of concepts to solve problems, not just the recollection and comprehension of basic facts (American Association for Advancement of Science [AAAS], 1989, 2011; National Research Council [NRC], 2003). Most recently, *Vision and Change* put forth the standard that undergraduates in biology develop competencies, including applying the process of science, using quantitative reasoning, and using modeling and simulation (AAAS, 2011). These competencies can be categorized as higher-order cognitive skills (HOCS; Zoller, 1993; Crowe *et al.*, 2008).

DOI: 10.1187/cbe.12-03-0024
Address correspondence to: P. P. Lemons (plemons@uga.edu).

Many faculty agree that students' development of HOCS is a primary objective of college, but data show that few college science courses actually teach or assess these skills. In a study from the Commission on Teacher Credentialing in California and the Center for Critical Thinking at Sonoma State University (Paul *et al.*, 1997), college and university faculty throughout California were surveyed to assess current teaching practices and knowledge of critical thinking, a construct that includes but is broader than HOCS. Of the faculty surveyed, 89% claimed critical thinking as an objective in their courses. Of the same faculty, only 19% could explain what critical thinking is, and only 9% were teaching for critical thinking (Paul *et al.*, 1997). Similarly, science, technology, engineering, and mathematics (STEM) courses and assignments often fail to promote the development of HOCS, as several studies have shown (Reynolds and Moskovitz, 2008; Momsen *et al.*, 2010; Ebert-May *et al.*, 2011).

One of the best ways to help students develop HOCS is to make HOCS questioning a regular part of their course work, because students' approaches to course work are strongly influenced by the type of questioning the instructor uses. For example, one study showed students expecting a multiple-choice exam focused their note-taking efforts on facts and details, whereas those expecting essay tests concentrated on main ideas (Nolen and Haladyna, 1990). Another study

documented that students are discouraged from trying to deeply understand material when exams only ask for memorization of large volumes of facts (Entwistle and Entwistle, 1992). Other researchers have shown that students tend to use deeper, more active approaches to studying when they are preparing for exams that include essay or HOCS questions (Traub and MacRury, 1990; Stanger-Hall, 2012). Clearly, using HOCS questions on exams and other assignments is an important strategy for teaching higher-order cognition.

Most faculty struggle to craft HOCS questions. A HOCS question typically includes a scenario that is novel to students and may also include graphs, figures, case studies, or research designs (Bissell and Lemons, 2006; Crowe *et al.*, 2008). Instructors who want to write HOCS questions from scratch end up spending a tremendous amount of time researching and writing a single question. Instructors who want to use test banks or modify questions starting with examples from their own experience or from colleagues face additional challenges. Biologists can easily find examples of questions requiring recollection and comprehension of facts, but examples of questions that also promote higher-order cognition are less plentiful.

Bloom's Taxonomy (Bloom *et al.*, 1956; Anderson and Krathwohl, 2001) has been used widely by biologists to overcome the challenges of writing HOCS questions (Hoste, 1982; Bissell and Lemons, 2006; Armstrong *et al.*, 2007; Freeman *et al.*, 2007; Stanger-Hall *et al.*, 2011). Bloom's Taxonomy helps instructors, because it provides descriptive vocabulary for higher-order learning objectives. Recently, Crowe and colleagues (2008) strengthened the connection between Bloom's Taxonomy and assessment in biology classrooms by providing the Blooming Biology Tool (BBT). The BBT helps biology instructors make use of Bloom's Taxonomy by providing general examples of biology questions that require different Bloom's skills, recommending types of questions that can be used for each Bloom's level, and articulating characteristics of multiple-choice questions for each Bloom's level.

Yet biologists still struggle to create good HOCS questions. Neither Bloom's Taxonomy nor the BBT describe all of the steps needed to construct a valid HOCS question, nor do they address some of the ideas about HOCS questioning that matter to biologists, such as question difficulty. Indeed, Anderson and Krathwohl (2001) and Crowe *et al.* (2008) called for additional research to create a more complete tool for linking higher-order learning objectives with valid HOCS questions.

In this study, we aimed to document biologists' ideas about HOCS questioning. We accomplished this by studying the conversations of nine biologists who prepared a set of ~40 HOCS questions to be used in introductory biology courses. In this paper, we report our qualitative analysis of the biologists' conversations about HOCS questioning. We also present a framework for additional research on biologists' conceptions and assumptions about HOCS questioning and discuss the implications for people who write HOCS questions or train others to write them.

METHODS

Context

Our study was based on the conversations of biologists who were writing or reviewing a set of ~40 HOCS questions. The biologists were preparing the HOCS questions for use in a study of the impact of higher-order clicker case studies compared with lower-order clicker case studies in introductory biology courses (Andrews *et al.*, 2012). Clicker case studies combine an engaging story, scientific content, and multiple-choice clicker questions (Herreid, 2007).

Table 1. Lower-order and higher-order cognitive skills

Cognitive skill level	Actions required by cognitive skill level
Lower order	Recall (memorize) facts, figures, and basic processes Know vocabulary and definitions Understand and illustrate information Includes Bloom's categories of knowledge and comprehension
Higher order	Use information, methods, concepts, or theories in new situations Predict consequences and outcomes Solve problems in which students must select the approach to use Break down a problem into its parts Identify the critical components of a new problem See patterns and organization of parts (e.g., classify, order) Determine the quality/importance of different pieces of information Discriminate among ideas Weigh the relative value of different pieces of evidence to determine the likelihood of certain outcomes/scenarios Make choices based on reasoned argument Includes Bloom's categories application, analysis, and evaluation

Higher-order clicker case studies were defined as clicker cases with at least 50% HOCS questions, whereas lower-order clicker cases were defined as clicker cases with all lower-order cognitive skills (LOCS) questions. Eight higher-order clicker case studies and eight matched lower-order clicker case studies were prepared on topics that included metabolism, DNA replication, Mendelian genetics, and evolution.

The writing of HOCS clicker questions was guided by a tool derived from Bloom's Taxonomy (Bloom *et al.*, 1956; Anderson and Krathwohl, 2001), the BBT (Crowe *et al.*, 2008), and related work (Bissell and Lemons, 2006). Specifically, this tool aligns action verbs with LOCS and HOCS (Table 1). Clicker questions primarily included multiple-choice questions with a single, correct answer.

A few of the clicker questions were multiple choice with more than one reasonable solution. For questions with more than one reasonable solution, students were asked to determine the relative importance of several pieces of evidence. Questions with more than one reasonable solution can encourage students to analyze evidence and use reasoning to decide why one piece of evidence is more important than another, even though students may not have the expertise to determine the actual correct answer. These are similar to the evaluation questions described in the BBT (Crowe *et al.*, 2008). Although questions with more than one good answer could be used in some testing settings, the few used in the case studies for this project were not intended to mimic test questions, but to create an opportunity for higher-order cognition.

To improve the validity and quality of HOCS questions, biologists who were not involved with question writing rated every question. Questions that did not rate as higher order were revised and rated again until they earned a higher-order rating.

Participants

The participants in this study included two teams of biologists. The Instructor Team worked with one investigator (P.P.L.) to write the HOCS questions and implemented the clicker cases in their courses for the study of case impact. The Rater Team rated the questions prepared by the Instructor Team.

The Instructor Team and P.P.L. worked together for 1 yr (June 2009 to May 2010) and met 10 times for about an hour per meeting. Meetings during the first 6 mo were primarily for the purpose of establishing processes for question writing and critiquing. Meetings during the second 6 mo were for the purpose of discussing how the case studies worked in the classroom and how the case studies should be revised before the data collection phase.

The Instructor Team consisted of two similar participants. Both taught introductory biology at a large, public, Research I institution. The classes they taught ranged in size from 100 to 350 students per semester. Their classes used clicker technology both for traditional lecture class periods and case study class periods. At the time of this study, the instructors had taught introductory biology for more than 10 yr each.

The Rater Team underwent intensive training to prepare to rate clicker questions. First, they reviewed a set of materials, including a description of the difference between LOCS and HOCS questions and an example clicker case study with several questions (both LOCS and HOCS). Second, they completed a 4-h training workshop, led by P.P.L, in which they discussed LOCS and HOCS questioning in general and analyzed and rated 26 clicker questions as a group. Raters then independently rated the remaining clicker questions. The Rater Team's work took place in February 2010.

The Rater Team consisted of seven participants from a large, public, Research I institution, including: four senior graduate students in the biological sciences, two professors in the biological sciences, and one lab coordinator in the biological sciences. At the time of this study, Rater Team participants who were graduate students had taught for a range of five to seven semesters, primarily as teaching assistants; Rater Team participants who were professors had taught for a range of 4–10 yr, primarily teaching upper-level courses for biology majors; the Rater Team participant who was a lab coordinator had more than 10 yr of teaching experience, primarily teaching introductory-level courses for nonmajors.

Data Collection

Two data sources were used to discover biologists' ideas about HOCS questioning. P.P.L. collected field notes while utilizing participant observation (Patton, 2002). Participant observation is a research strategy utilized to uncover patterns of thought within small groups in which the researcher is both participant and observer, actively and personally engaging in the process, while simultaneously making careful observations and writing detailed field notes (Dewalt and Dewalt, 2011). Also, transcripts were developed from audio recordings of Instructor and Rater Team meetings and utilized as a source of data.

P.P.L. took field notes based on her participant observation during the first seven meetings of the Instructor Team. She transcribed audio recordings from the remaining three meetings with the Instructor Team and all of the meetings with the Rater Team. Field notes and transcripts were entered into MaxQDA (VERBI GmbH, Berlin, Germany), a software package for qualitative data analysis.

Data were collected under exempt status at the University of Georgia (project #2011-10909-0).

Data Analysis

We used qualitative analysis of field notes and transcripts from Instructor and Rater Team meetings to document biologists' ideas about HOCS questioning. Our qualitative analytical method aligned with grounded theory (Creswell, 2007; Glaser and Strauss, 2010), a method that considers "What theory emerges from systematic comparative analysis and is grounded in fieldwork so as to explain what has been and is observed?" (Patton, 2002). Recently, educational researchers in biology have used grounded theory analysis to study science faculty with education specialties (Bush *et al.*, 2011) and biology undergraduates' misconceptions about genetic drift (Andrews *et al.*, 2012).

Rigor in qualitative research has been defined as the attempt to make "data and explanatory schemes as public and replicable as possible" (Denzen, 1978, p. 7, quoted in Anfara *et al.*, 2002). Therefore, to ensure the rigor of our analysis, both authors (P.P.L. and J.D.L.) followed the systematic approach described here. We independently read and reread meeting field notes and transcripts, coding patterns of thought expressed by participants. Then we met and discussed our coding themes. We critically analyzed each other's coding using peer examination (Merriam, 2009). We coded field notes and transcripts by placing excerpts in thematic categories if they met two criteria: 1) participants used particular words, and 2) the context of the conversation made it clear that participants' use of these words was meaningful and purposeful. For example, excerpts in the category Difficulty 1) include the words difficult, challenging, easy, etc., and 2) are clearly focused on the idea of determining how challenging a question is for students. We marked the beginning of a coded excerpt when the conversation turned toward a particular category and marked the end of a coded excerpt when the conversation turned away from that category. Coded excerpts ranged in size from one statement by a single participant to a conversation of up to 21 statements by different participants, but on average coded excerpts were about three statements in length. Through peer examination, we confirmed the presence of the following categories within the data: Bloom's, Difficulty, Time Required, Student Experience, and Correct Answers.

Reliability of the Data

We addressed two potential concerns about data reliability in our analysis.

First, to protect from bias by the participant observer (P.P.L.), field notes were used only as a method of documenting the initial occurrence of themes (see Supplemental Material for examples of field notes). In recording the frequency of thematic categories from the field notes, we took care not to overrepresent their presence. If a thematic category emerged in a set of field notes, we coded it only one time. Additionally, we only report thematic categories from field notes that were confirmed with transcripts from later meetings. All thematic categories first emerged in the field notes except one, Correct Answers.

Second, to address interrater reliability, we sought to build consensus throughout the analysis phase, using independent coding followed by peer examination. We report themes only from data for which we reached 100% consensus.

RESULTS

We describe here the thematic categories, illustrating each category with quotes from coded excerpts. When participants refer to clicker questions, the questions are either cited (for previously published questions) or presented as figures (for unpublished questions). In the interest of brevity, we quote only the most salient sections of a coded excerpt, rather than the entire coded excerpt. Quotes from the Instructor Team are labeled I1 or I2, for Instructor 1 or Instructor 2. Quotes from the Rater Team are labeled R1, R2, etc., for Rater 1, Rater 2, etc.

The Use of Bloom's Taxonomy

Multiple times within the data, participants referred to the derivative of Bloom's Taxonomy they used for question writing and review (Table 1). For example, the Instructor Team constructed a question asking students to determine the best placement of mitochondria on a phylogenetic tree (Brickman, 2008, slide 40). They compared the question with Table 1 and noted the question requires students to use information in a new situation. Students were to be given facts about the similarities between mitochondria/chloroplasts and bacterial cells during the case study, which is evidence that mitochondria/chloroplasts evolved from a bacterial ancestor. The Instructor Team designed the mitochondria question to prompt for application of this knowledge by placing mitochondria on the tree.

The Rater Team also relied on Table 1, literally looking back and forth from the language of the question to the language of Table 1 in order to rate questions. For example, when discussing a question on genetic engineering (Figure 1), one Rater Team participant (R1) said, "I'm totally on the fence right now, but I keep going back to this table [Table 1], and they are using this information in a new setting. They are identifying critical components of a new problem. . .they are determining the quality of importance. . .I mean they are doing a lot of these things [e.g., using information in new situations]. . ."

Similarly, when discussing a question about mutation (Figure 2), one Rater Team participant (R1) commented: "On the list [Table 1], you're breaking down the problem into parts and you're predicting consequences."

Clicker Question: Which of the following experiments would NOT work to produce a male mouse through genetic engineering. You start with a normal mouse embryo with:

A. the normal number of autosomes and two X chromosomes and then add a Y chromosome to the cells of the embryo.
B. the normal number of autosomes and two X chromosomes then add an SRY gene to the cells of the embryo.
C. the normal number of autosomes, an X chromosome, and a Y chromosome, then add nothing to the cells of the embryo.
D. the normal number of autosomes and two X chromosomes, then remove one of the X chromosomes from the cells of the embryo.

Figure 1. Genetic engineering question from a case study on sex determination, chromosomal crossing over, and sex linkage.

It's Not Just Bloom's

Although 37 distinct excerpts of data were coded as Bloom's (five in field notes, 32 in transcripts), 60 distinct excerpts of data did not fit into the Bloom's category (six in field notes, 54 in transcripts). We placed these excerpts in other categories that emerged from our analysis.

For example, when participants worked on a question asking students to choose a primer for a given DNA sequence (Armstrong *et al.*, 2009, slide 19), they referenced Bloom's (Table 1), but they also referenced other ideas about higher-order questioning. The Instructor Team called the question higher order and placed it in the case so that students would encounter it right after they were given information about complementary base pairing, DNA replication, and polymerase chain reaction (PCR). They designed the question to prompt for application of information about PCR (Table 1). When the Rater Team evaluated the question, there was some discussion about whether the question would elicit recollection of the basic process of PCR or application of PCR. However, most of the discussion was about other ideas, such as the difficulty of the question and whether students would have previous experience working a similar problem. Consider this excerpt:

> R1: "In going back and looking at those slides [slides leading up to the question], I feel like they were given

Clicker Question: Which figure depicts what the DNA would look like if DNA polymerase does not function in a cell replicating its DNA? Be prepared to explain the figure you chose and also why you did not choose the other figures.

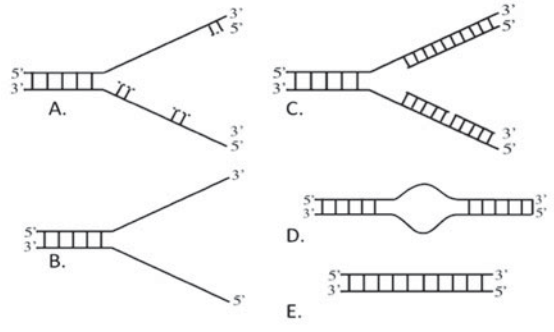


Figure 2. Mutation question from a case study on DNA replication and PCR.

all the information they need to answer this question, it is relatively straightforward. I am looking at this list. . .in the higher order, I don't see them doing any of these things."

R4: "It's just a more difficult lower-order question."

R2: "To me, I agree with it being a higher. . .because of the 5′→ 3′. If that 5′→ 3′ wasn't there. . .I would consider it a low, but because they have to think about the direction of the primer and how to find it in the answer that's making them think a little bit more. . ."

[After a bit more discussion. . .]

R3: "I still think it is lower order but. . .I agree it is a more difficult lower order. . .'cause all they need to go back and. . .really understood what was taught. . .they just need to go to retrieve it in a correct way. . .it's not something that they have to really work through in step, step, step in order to get to the answer. . ."

R5: "But if you've never had this before, and the first exposure you have is those few slides ago. . ."

R3: "Except the slide before that was talking about it. . ."

R2: "But it's still teaching. They haven't been able to apply; they are just getting knowledge in their heads, so this is the time they are applying it. And the 5′→ 3′ might be a little difficult. *Might* be a little difficult to them."

All Rater Team participants acknowledged that students would see pertinent content prior to the question (Armstrong *et al.*, 2009, slide 19). Despite this agreement, participants who believed students would easily find the answer to the question rated it lower order, and participants who believed students would be challenged by the question rated it higher order. Rater Team members never reached consensus on the question.

For elements of questioning that could not be characterized using Bloom's, we identified four thematic categories revealing participants' conceptions of higher-order questioning: Difficulty, Time Required, Student Experience, and Correct Answers. These categories are summarized in Table 2 and fully described below. We also offer our interpretation of how each category relates to the process of writing higher-order questions.

Category 1: Difficulty

The Difficulty category reflects participants' thinking about whether a question was expected to be challenging for students, whether the concepts referenced in a question are generally hard to apply, and how students typically perform on similar questions.

Almost all participants used the words "difficult," "challenging," or "easy" to describe why particular questions did or did not require higher-order cognition. The crux of the matter for some questions was not Bloom's (Table 1), but the perceived level of challenge the question would provide to students.

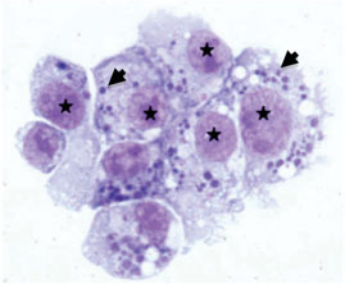
The Instructor Team expressed ideas about difficulty throughout their work. During their very first meeting, during which they developed methods to define higher-order questions, they discussed the difference between cognitive level and difficulty and noted that a difficult question is not necessarily a higher-order question. When the Instructor Team tested case studies on students and found the questions were easy (i.e., most students got the questions right), they worried the questions might not be higher order, despite having used Table 1 as a guide in the design process. Additionally, they judged the merit of questions on the

Table 2. Thematic categories with brief descriptions and representative quotes

Category	This category reflects participants' thoughts about:	Quote
Difficulty	Whether the question is expected to be challenging for students; whether the concepts used in the question are hard to apply; how students typically perform on similar questions. (4, 27)	Rater Team participant: "Yeah, it's just more difficult. . .The reason I think it's higher order is I think this concept is sort of hard to grasp unless they've thought about this before. . ."
Time Required	How long it takes to arrive at an answer to the question. (1, 6)	Rater Team participant: "The reason I think it's lower is because you can go through it very quickly. You can see that you have one differential pair at the bottom. You can quickly rule out female bee. . .so it's very quick without much thinking."
Student Experience	How experienced students are in solving similar questions; whether the question is routine; whether the question requires students to use a new or well-practiced approach. (1, 9)	Rater Team participant: "I thought of higher [order] because I had to decide, how am I going to tackle this question from everything that I know?"
Correct Answers	The merit of questions intentionally designed to have more than one reasonable solution, particularly multiple-choice questions. (0, 12)	Instructor Team participant: "With multiple choice there is a correct answer. For questions where there isn't a correct answer, I'd much rather go through it in the class. . .and say what do you think are some of the changes or differences. . .rather than have a multiple-choice question."

Biologists who participated in this study were concerned with several dimensions of questioning, in addition to Bloom's-like definitions. Through qualitative analysis, their concerns were categorized as *Difficulty*, *Time required*, *Student Experience*, and *Correct answers*. The number of occurrences of each category is shown in parentheses after the category description, with the first and second numbers representing occurrences in field notes and transcripts, respectively.

Clicker Question: The following photomicrograph shows several human cells stained as large light purple structures with purple nuclei (stars) infected with a pathogenic organism shown as small dark purple circles (arrowhead). These pathogens are most likely a(n)



A.Virus

B.Prokaryote

C.Eukaryote

Figure 3. Pathogenic organism question from a case study on cell structure.

basis of student performance. Even though the Instructor Team was regularly stating that cognitive level and difficulty may be two different things, they made a practice of judging the cognitive level of a question *based on* difficulty. For example, the Instructor Team discussed a question that asked students to identify a pathogenic organism (Figure 3). When the Instructor Team put this question into the case, they instinctively called it higher order, because previous students had typically answered it incorrectly. They were surprised to learn the Rater Team did not agree the question was higher order:

> I1: "I don't know why, but my students did much better on this question than they have previously, at least [that is] my perception."
>
> I2: "Mine said 50% virus still, 34% said prokaryote [the correct answer], and 15% said eukaryote. . ."
>
> P.P.L.: "Nobody [among the Rater Team] thought that was a higher-order question."
>
> I1: "Really?"
>
> I2 [clarifying the rationale]: "Just the understanding of what the size is of these?"
>
> P.P.L.: "To me, it's an application."
>
> I1: "They are trying to gather visual information with other logical explanations."

The Rater Team engaged in similar discussions about Difficulty (see Table 2), especially with questions with controversial ratings. One Rater Team participant repeatedly argued that one controversial question was higher order, because it was "difficult," and that another controversial question was *not* higher-order, because it was "not extra hard."

Participants did not think all difficult questions were higher-order questions. Almost all participants separated difficulty and cognitive level in their statements. Some participants even gave questions two labels, such as calling them "difficult lower-order questions." In contrast, participants indicated that easy questions were by definition not higher-order questions. As a result, participants tried to use two criteria for evaluating questions: 1) Does the question either explicitly or implicitly prompt for cognitive steps like those described as Higher order (Bloom's category) in Table 1? 2) Is the question difficult for students?

Category 2: Time Required

Closely related to Difficulty is the category Time Required. The Time Required category shows participants' thinking about the interaction between question length and cognitive level. Participants repeatedly used words like "time" to explain their ideas about the differences between lower-order and higher-order questions. Their general ideas were that students could answer lower-order questions quickly and that higher-order questions were time-consuming to answer. For example, the Instructor Team worried about the impact of question length on the study of higher-order clicker case studies compared with lower-order clicker case studies. They were afraid the higher-order case studies would take up a lot more class time than lower-order case studies; they had observed that good higher-order questions require more time to solve. After testing the case studies in the classroom, one Instructor Team participant noted that class discussions of lower-order questions did not take very much time; there was not a lot for students to say, because they arrived at a correct answer quickly.

Another illustration of Time Required comes from the Rater Team's discussion of the primer question previously referenced (Armstrong *et al.*, 2009, slide 19). The Rater Team could not agree on the rating for this question, because they disagreed about its level of difficulty (category: Difficulty). In an effort to resolve the disagreement and inform the rating of additional questions, the Rater Team tried to come up with ways to make the primer question a valid higher-order question. One participant suggested revising the question by giving students a primer sequence and asking them to produce a PCR product. Other participants commented on this suggestion, noting that it would take a lot more time and might or might not accomplish anything different cognitively:

> R2: "You could switch the question to a little title like, 'You need to amplify this. . .this is your primer sequence. What would be the PCR product?'"
>
> R5: "Uh. . .I think it's going to take more time. You will have to give them a fair amount of time to resolve it, and I'm not sure it's really getting much more information. It's more labor intensive for them."
>
> R1: "I think it makes it harder [than] to just look at the five selections and immediately go, 'Oh, that must be it,' because right now [in the current form of the question]. . .a fairly bright student wouldn't really have to think about it, and they could just say, 'Oh, clearly, it's. . .'"

A different participant suggested revising the question to make the primer sequence match the opposite end (right end) of the given sequence. Participants thought this revision would be more time-consuming and difficult:

> R5: "You have to have a little more thought go into it if the primer is coming from the other end, because you actually have to flip it and think about [it], and that would really test whether you really understand what you're talking about."

Clicker Question: In a Drosophila experiment, a cross was made between homozygous wild-type females and yellow-bodied males. All the resulting F_1 offspring were wild type. However, the adult flies of the F_2 generation (resulting from matings of the F_1s) had these characteristics:

Sex	Phenotype	#
male	wild	123
male	yellow	116
female	wild	240

The mutant allele for yellow body is:

A. Recessive, X-linked
B. Recessive, not X-linked
C. Dominant, X-linked
D. Dominant, not X-linked
E. None of the above

Figure 4. Genetics question from a case study on sex determination, chromosomal crossing over, and sex linkage.

> R4: "But that's still knowledge. That's what we're saying. That would be the same level, but it is just more difficult and time-consuming."

For the primer question, when participants tried to think of ways to make it more difficult (which in most of their minds would make it a higher-order question), their solutions involved making the question take more time.

We interpret these data to mean that Time Required is a metric used to gauge question difficulty. In our data set Time Required almost always occurred in conversations that were also coded with Difficulty, and participants seemed to be assessing question difficulty based on whether the question could be answered quickly. In fact, participants not only thought about how many seconds or minutes a question would take, but they also thought about how many cognitive steps a student would go through to solve a question.

Category 3: Student Experience

The Student Experience category indicates participants' thinking about student familiarity solving similar questions, whether a question is routine, and whether a question requires students to use a new or well-practiced approach. Generally speaking, participants did not feel comfortable calling a question higher order if they expected students to be experienced with the question type. They feared that routine questions would be "plug-and-chug" questions instead of thought-provoking questions. Rather, participants said that higher-order questions should prompt students to use new ways of thinking.

One example comes from a genetics question that could be solved using a Punnett square (Figure 4). When Rater Team participants discussed the question, most participants thought it was higher order, but one participant was reluctant to call it higher order, because of the use of a Punnett square. Punnett squares were to be taught prior to this particular case, and as one reviewer noted, Punnett squares are something students "do a lot":

> R7: "I'm hedging toward lower. I'm sort of going between [lower and higher], because I feel like the lower-order part is doing the Punnett square. The process of doing the Punnett square."

> P.P.L.: "And you think it's lower order because they've done it before?"

> R7: "Because they've done it before and. . .they're doing another Punnett square. . ."

> [Reviewer 7 persisted until Reviewer 2 pointed out that the question prompted her to select an approach for solving the problem.]

> R2: "I thought of higher because I had to decide, 'How am I going to tackle this question from everything that I know?'"

Once Reviewer 2 made this statement, other reviewers noted that, even though the question involved a Punnett square, it also required nonroutine skills, such as determining which pieces of information among the data about sex and phenotype were most important. This line of reasoning was convincing to Reviewer 7, who eventually agreed that the question required a new way of thinking about genetics problems.

These data suggest that Student Experience with the particular content and structure of the question is an important indicator of the difficulty of a question. Student Experience sometimes occurred in conversations also coded with Difficulty. In these instances, participants appeared to be gauging difficulty based on their perception of the level of student experience with that problem type.

Category 4: Correct Answers

Correct Answers is the final category that emerged from our data and reflects participants' thinking about the merit of questions with multiple reasonable solutions. Some participants thought questions with multiple reasonable solutions were inappropriate when formatted as multiple-choice questions, but other participants thought they were valuable in promoting higher-order cognition.

For example, consider the jaw question shown in Figure 5. An expert biological anthropologist would know that all of the evidence shown, (i.e., incisors, canines, premolars and molars, jaw shape, and spacing among teeth) inform a researcher whether the jaw is ape-like or human-like (Choice 5), but students were not expected to know this. Rather, prior to this slide, students were asked to examine pictures of chimpanzee, *Australopithecus afarensis*, and *Homo*

Clicker Question: Here is a jaw like the ones you've studied on the previous two slides. How informative are the noted features in determining whether this jaw is from an ape or a human/ancestor to modern human? Categorize the noted features as informative or uninformative. Be prepared to explain your rationale.

1. Informative: A; Uninformative: B, C, D, E.
2. Informative: A, B; Uninformative: C, D, E.
3. Informative: A, B, C; Uninformative: D, E
4. Informative: A,B,E; Uninformative: C, D
5. All features are informative.
6. I chose another categorization.

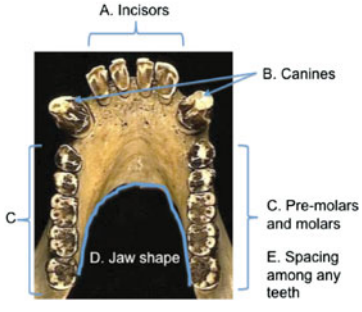


Figure 5. Jaw analysis question from a case study on human evolution.

sapiens jaws and were asked to note the similarities and differences among them. Students were given the question in Figure 5 without any additional instruction on the differences and were expected to notice that some pieces of evidence (e.g., the incisors) varied more among chimpanzees, *A. afarensis*, and *H. sapiens* than other pieces of evidence (e.g., premolars and molars). In this sense, the question could have more than one reasonable answer, because novice biology students using higher-order cognition might reason that answers 2, 3, 4, 5, or 6 are correct. Arguably, only answer 1 is not defensible from the perspective of a novice biologist.

When the Instructor Team discussed this question, Instructor 1 objected to the question, noting the idea that multiple-choice questions should have a single, correct answer. Instructor 2, on the other hand, thought the question forced students to commit to one good answer and sparked a lot of classroom discussion:

> I1: "I just generally found question 5 [Figure 5] in the format of the answers to be just too complicated. Just keeping track of which is informative and uninformative. I think we talked about it before. For a question like this, I'd like it a lot better as a discussion question, and this is what I'm thinking. I think we're going back to a previous case where there is an indefinite answer for multiple choice. Maybe just from my own background, it really goes against my grain. With multiple choice, there is a correct answer. For questions where there isn't a complete answer, I'd much rather go through it in the class. Have the class discuss, show different jaws, and say what do you think are some of the changes or differences do you think you can use to distinguish them apart rather than have a multiple choice question."
>
> I2: "Everyone has to commit. The problem with the discussion is you don't get anyone to commit. It [the multiple-choice question] forces them to commit."
>
> I1: "Could we make the question a little less complicated?"
>
> I2: "Most of mine did not want to choose. They did not want the jaw shape changes. They liked the canines absolutely. They almost wanted to order them in terms of which were the most valuable and which weren't. That seemed easier to them than this uninformative/informative thing. They wanted to order them. That's why they didn't like any of these. They were like, 'Well I thought the canines and the incisors. . .'"
>
> I1: "So, they liked the canines and spacing probably. . ."
>
> I2: "The only one they didn't want was the premolars. They wanted uninformative with just C. They did like D, so they were like, I like all of them but C, but that was not an option. It definitely sparked some conversation. I had no issues with it in that sense."

Another example of a question with more than one reasonable answer is shown in Figure 6. This question was also written with the expectation that students could arrive at multiple correct answers. In fact, although some expert biologists may choose answer A, others would disagree and point out that gender determination is a controversial topic. Instead, the expectation was for students to use their knowledge of variance in each of these traits—among males and females—to decide which pieces of evidence provide strong evidence that Santhi is female and which pieces provide weak or no evidence that Santhi is female.

Clicker Question: Imagine you are a member of the committee assigned to determine whether Santhi is female. Here are possible results (we don't know the real results):

A. Female genitalia: Yes

B. Breasts: Yes

C. Pubic hair: Yes

D. Regular menstrual cycle: Never

Rank these pieces of evidence in determining Santhi's sex from most (1) to least (4) important. Be prepared to explain your rationale.

A. 1/A, 2/B, 3/D, 4/C
B. 1/C, 2/A, 3/D, 4/B
C. 1/D, 2/A, 3/B, 4/C
D. 1/A, 2/D, 3/C, 4/B

Figure 6. Sex determination question from a case study on sex determination, chromosomal crossing over, and sex linkage.

The Rater Team discussed this question at length. When Rater Team participants initially looked at the question, some called it a case advancement question. Case advancement questions are questions that advance the story or theme but do not have a cognitive aim, per se. Rather, they serve the purpose of engaging or surveying students:

> R4: "This almost seems like a case advancement."
>
> R1: "Yeah, it does, like. . .there's not a right answer."
>
> R6: "That's what I was thinking."
>
> R1: "It's an opinion. . ."

At this point in the conversation, another Rater Team participant joined in, but not to state her ranking of the question. Rather, she was intrigued by a piece of evidence in the question—the fact that Santhi had never had a menstrual cycle (see Figure 6). Other Rater Team participants joined in the conversation expressing their concerns that students would not have not been told this fact in the case and, thus, may have varying amounts of knowledge about the impact of athletic activity on menstruation:

> R5: "With D [showing data about menstruation], you have to know something in advance which is. . .your body fat gets below a certain rate then you'd stop menstruating."
>
> R4: "I think different people might have, yeah, different, knowledge about that."
>
> R5: "So if you have that knowledge from somewhere else. . ."
>
> R3: "Yeah, but given what's been given, you know, the slides prior to this, it doesn't seem like these are things that they necessarily would have to. . .I mean especially when you are talking about most important versus least important."
>
> R4: "Yeah, exactly, *opinion*. . ."
>
> R5: "But I mean with athletes when you get low. . .I think 15% [body fat]. . .they stop menstruating. . ."

Later in the same discussion another Rater Team participant explained why she thought this was a case advancement question. She had come up with a solution that was not among the choices provided, so even though the question prompted students to execute higher-order cognition,

the ambiguity of the answer choices disqualified the question as higher-order:

> R1: "The order [answer] I would have chosen is not one of the options."
>
> P.P.L.: "OK, so what is the order [answer] you would have chosen?"
>
> R1: "I would have chosen A, D, B, C. . ."
>
> P.P.L.: ". . .if that were one of the options, would that have affected your rating of the question?"
>
> R1: "I still think it's probably case advancement. I understand what you were saying [addressing a participant who rated the question higher order] about 'making a determination,' but to me, these are too gray, and there's not one right answer. . ."

In the end, five of seven Rater Team participants rated this question as case advancement. They determined the question was ineligible for higher-order status, because all students would not have had equal exposure to the information necessary to evaluate the evidence and because the question might have more than one good answer *or* an answer that was not among the possible selections.

Most of the participants felt that questions with multiple reasonable solutions were not higher-order questions, even though they themselves used higher-order cognitive skills to solve them. Most of the conversation about questions with more than one good answer concerned the evidence presented. That is, the Instructor and Rater Teams spent a lot of time engaged in the data from the questions. They explained why they thought one piece of evidence was more weighty than another, debated which pieces of information were actually informative, and expressed frustration when they thought two answers might be equally good or when the answer they thought was best was not among the choices. These data show that biologists make assumptions about the role of questions in the classroom and that these assumptions powerfully influence the types of questioning they are willing to use.

DISCUSSION

These results confirm the utility of Bloom's Taxonomy–derived tools for biologists who attempt to write and evaluate higher-order questions. Both the Instructor and Rater Teams referenced a Bloom's-derived tool (Table 1) throughout their work, just as other biologists have used Bloom's in question writing and evaluation (Hoste 1982; Bissell and Lemons 2006; Armstrong *et al.*, 2007; Freeman *et al.*, 2007; Stanger-Hall *et al.*, 2011). Our results also suggest that biologists have ideas about higher-order questioning that go beyond Bloom's. About 62% of our categorical data (60 coded excerpts) did not fit into the Bloom's category but instead were categorized as Difficulty, Time Required, Student Experience, or Correct Answers.

Framework

Consistent with a grounded theory approach, the aim of our study was to generate a framework for further research that derives from data. Figure 7 depicts our framework, which provides a model of participants' ideas about higher-order questioning and how those ideas may relate to one another. It also includes hypotheses than can be explored with additional research. We placed each analytical category underneath the box Questions Intended to Assess HOCS to show that each category influenced biologists as they evaluated HOCS questions. Within the framework, we clustered categories and named the clusters. We clustered Bloom's, Difficulty, Time Required, and Student Experience and called that cluster Factors Used when Evaluating Potential HOCS Questions. We did this because participants used the language from Bloom's (Table 1), and they also attempted to analyze difficulty, time required, and student experience for each question. They used their analysis to judge the higher-order quality of a question.

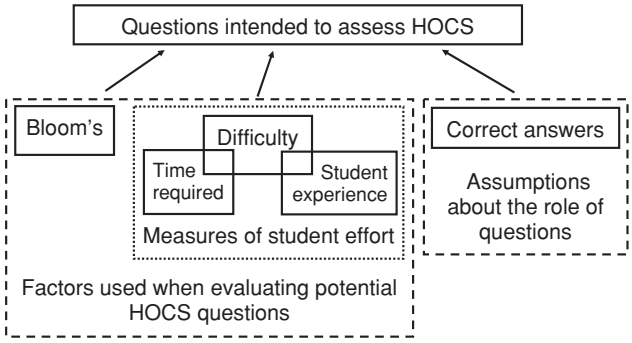


Figure 7. Framework for further research showing how biologists conceptualize questions intended to assess HOCS. The four thematic categories uncovered have been clustered to model how they may be related to each other. Two hypotheses emerge from this model: 1) Biologists intuitively look for ways to measure difficulty when evaluating potential HOCS questions, and time required and student experience are two of the ways they attempt to measure it. 2) Biologists' assumptions about the role of questions in student learning influence the types of HOCS questions they write and even their comfort with HOCS questioning.

In our framework, we created a cluster within the Factors cluster called Measures of Student Effort. In evaluating Difficulty, Time Required, and Student Experience, participants seemed to be looking for ways to describe the amount of student effort needed for a question. Within the Measures of Student Effort cluster, we placed Difficulty at the apex. Participants' biggest concern, next to Bloom's, was question difficulty. Participants did not say that all difficult questions are higher order; rather they said that a higher-order question is not easy. We also included Time Required and Student Experience in the Student Effort cluster, and we placed them overlapping Difficulty. There are three reasons for this arrangement. First, within our data set, Difficulty and Time Required and Difficulty and Student Experience were sometimes found in the same conversations. Second, looking across the entire data set and the context of each coded excerpt, we noted that Time Required was a way for participants to judge question difficulty, as was Student Experience. Third, this grouping is consistent with theory and empirical research from educational psychology. In educational psychology, a standard measure for the difficulty of problems is the amount of time needed to complete the problem (e.g., Kotovsky *et al.*, 1985; Sweller and Chandler, 1994). Also, based on cognitive load theory, which provides a framework for thinking

about the difficulty of learning information, student experience can be thought of as the extent to which students are expected to have schemas for a particular question (Sweller and Chandler, 1994). If students have a schema for a particular type of question, the question may be speedy and straightforward to address. If they do not, the problem may be time-consuming and difficult.

In our framework, we placed Correct Answers in a cluster by itself and called the cluster Assumptions about the Role of Questions. Participants' conversations about correct answers reflected their assumptions about whether a multiple-choice question should have multiple, reasonable answers; whether students should be expected to answer a question when they've not been told what the answer is; and whether ambiguity in questioning is acceptable. We propose that Correct Answers is only one of several assumptions biologists may make about the role of questions in the biology classroom.

Two hypotheses emerge from our framework. First, we hypothesize that biologists intuitively look for ways to measure Difficulty when attempting to write higher-order questions, with Time Required and Student Experience as two measures they attempt to use. On the basis of this hypothesis, we predict that a broader population of biologists, beyond those in our sample, would also consider Difficulty, Time Required, and Student Experience when they evaluate potential higher-order questions. In informal conversations with biologists prior to this study, we have noticed that difficulty is often discussed. Additionally, another study, focused on K–12 item writers, showed that item writers conflated cognitive level and difficulty (Wyse and Viger, 2011). We suspect that our prediction would bear out in further studies.

More importantly, however, our first hypothesis raises a question. That is, given current knowledge from educational research, are the biologists right? Are higher-order questions more difficult for students than lower-order questions? The data are equivocal. Some research has shown that students do not perform as well on exams with higher Bloom's levels (Knecht, 2001; Freeman and Parks, 2010). Additionally, Freeman and colleagues (2011), found a strong negative association between a weighted Bloom's index and predicted exam score. In contrast, Momsen and colleagues found no relationship between the difficulty of biology test items (measured by the percentage of students who answered the item correctly) and the Bloom's level of those items (J.L. Momsen, personal communication). The discrepancy of these findings suggests further research is needed to fully describe the relationship between higher-order questions and item difficulty. This research should include studies that correlate Bloom's ratings of items with difficulty as defined by psychometrics, including classical item analysis, as well as item response theory (Murphy and Davidshofer, 2005). Another approach to understanding the relationship between Bloom's ratings and item difficulty is to investigate *how* students answer questions by asking them to document their solutions to questions of various Bloom's ratings. This method reveals whether questions thought to be higher order actually prompt for higher-order processes by students (unpublished observations). Regardless of the findings from this research, however, Bloom's ratings and item difficulty are technically two different concepts, and even if they positively correlate, biologists should be discouraged from conflating them.

Second, we hypothesize that biologists' assumptions about the role of questions, such as the assumption that multiple-choice questions should have a single correct answer, strongly influence the types of higher-order questions they write and even their comfort with higher-order questioning. To explore this hypothesis, we intend to continue studying instructors who are engaged in higher-order question writing. We would like to probe to discover more of their assumptions, as the assumptions could help to explain why higher-order questioning is not more prevalent in biology classes, when biologists claim they value it. Indeed, recent research on change in STEM undergraduate education shows that changing the practices of STEM educators requires not only dissemination of curricula and techniques (like Bloom's) but also attention to educators' beliefs about teaching and learning (Henderson, 2005; Henderson *et al.*, 2011).

Implications for Teaching and Professional Development

Our data suggest that even biologists with experience teaching and using Bloom's Taxonomy would benefit from more guidance about higher-order questioning. Biologists who learn about Bloom's Taxonomy are not blank slates. If asked to write or evaluate higher-order questions, they do not use Bloom's Taxonomy in a vacuum. Rather, they bring to the task their own conceptions and beliefs about higher-order questioning. Some of these conceptions are misguided, and further training could help correct these misconceptions. For example, biologists who use Bloom's can be taught that quantitative methods exist for precisely determining item difficulty, including classical item analysis and item response theory (Murphy and Davidshofer, 2005); that item difficulty is an important and measurable characteristic of a question *distinct* from its Bloom's rating; and that more research is needed regarding whether higher-order questions are more difficult questions. Of course, any additions to Bloom's Taxonomy should be considered with great care, lest the simplicity of the taxonomy be lost.

Our data also suggest some biologists may mistakenly think that providing students with opportunities to practice higher-order questions gives too much away. That is, some biologists may think they should save all the higher-order questions for high-stakes exams, so the questions will be higher order, hard, time-consuming, and new to students. On the contrary, making the classroom a place in which students practice science is the idea behind many national calls for reform in undergraduate science classrooms (AAAS, 1989, 2011; NRC, 2003; Handelsman *et al.*, 2004), and many published reports show how this can be done (e.g., Bogucka and Wood, 2009; Freeman *et al.*, 2011; Hoskins *et al.*, 2011). It stands to reason that specific higher-order questions may become more routine when students practice them. For example, if students practice reading phylogenetic trees and receive feedback about their performance, more of them should respond correctly to a tree-reading question than if the same students were given no practice and no feedback. But research in engineering education shows students need thousands of hours of deliberate practice to achieve broad expertise in applying knowledge and skills to solve problems like experts (Litzinger *et al.*, 2011).

For these reasons, we propose that biologists give their students lots of practice with higher-order questions. The

experience may make some questions easier, quicker, and more routine than they would have been without the practice, but the experience also may build expertise, enabling students to solve a greater diversity of higher-order questions than they could have solved without practice. This end is precisely the goal of using HOCS in the first place. Furthermore, under these conditions, students may perceive courses as simultaneously supportive and academically challenging, characteristics that make a difference in graduation rates, student educational gains, and student satisfaction with college (Laird *et al.*, 2008a, 2008b).

Our work sheds light on biologists' conceptions and assumptions about higher-order questions. Not all biologists have a deep understanding of educational psychology or research, but most biologists understand biology as a discipline and have intuitions about what it means to solve problems in biology. Our work suggests professional development personnel have an opportunity to profoundly influence undergraduate biology classrooms by recognizing and respecting the ideas instructors bring to the HOCS question-writing process. Efforts to change the undergraduate biology classroom that do not consider biologists' ideas are likely to fail to gain the traction needed for sustained change in education in the biology community at large.

ACKNOWLEDGMENTS

The context for this research was supported by the NSF under Grant No. DUE-0920264 awarded to the National Center for Case Study Teaching in Science. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. We thank Clyde F. Herreid and David Terry for the opportunity to use higher-order, clicker case construction as a context for our research. We thank all the participants in the study. We also thank members of The University of Georgia Science Education Research Group, Luanna Prevost, Michelle Smith, William B. Wood, and our reviewers for their insights on the manuscript. Norris Armstrong wrote the question shown in Figure 2. Peggy Brickman wrote the questions shown in Figures 3 and 4.

REFERENCES

American Association for the Advancement of Science (AAAS) (1989). Science for All Americans, Washington, DC.

AAAS (2011). Vision and Change in Undergraduate Biology Education: A Call to Action, Washington, DC.

Anderson LW, Krathwohl D (eds.) (2001). A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives, New York: Longman.

Andrews TM, Price RM, Mead LS, McElhinny TL, Thanukos A, Perez KE, Herreid CF, Terry D, Lemons PP (2012). Biology undergraduates' misconceptions about genetic drift. CBE Life Sci Educ *11*, 248–259.

Anfara VA, Jr., Brown KM, Mangione TL (2002). Qualitative analysis on stage: making the research process more public. Educ Res *31*, 28–38.

Armstrong N, Chang S, Brickman M (2007). Cooperative learning in industrial-sized biology classes. CBE Life Sci Educ *6*, 163–171.

Armstrong N, Platt T, Brickman P (2009). The Case of the Druid Dracula: Clicker Case Version, Buffalo, NY: National Center for Case Study Teaching in Science Case Collection. http://sciencecases.lib.buffalo.edu/cs/collection/detail.asp?case_id=493&id=493 (accessed 1 September 2012).

Bissell A, Lemons PP (2006). A new method for assessing critical thinking in the classroom. BioScience *56*, 66–72.

Bloom BS, Englehart MB, Furst EJ, Hill WH, Krathwohl DR (1956). Taxonomy of Educational Objectives: The Classification of Educational Goals, New York: McKay.

Bogucka R, Wood E (2009). How to read scientific research articles: a hands-on classroom exercise. Iss Sci Technol Librarianship *Fall 2009*, doi:10.5062/F4S180FF.

Brickman P (2008). Take Two and Call Me in the Morning: A Case Study in Cell Structure and Function, Buffalo, NY: National Center for Case Study Teaching in Science Case Collection. http://sciencecases.lib.buffalo.edu/cs/collection/detail.asp?case_id=500&id=500 (accessed 22 September 2012).

Bush SD, Pelaez NJ, Rudd JA, Stevens MT, Tanner KD, Williams KS (2011). Investigation of science faculty with education specialities within the largest university system in the United States. CBE Life Sci Educ *10*, 25–42.

Creswell JW (2007). Qualitative Inquiry and Research Design: Choosing among Five Approaches, 2nd ed., Thousand Oaks, CA: Sage.

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*, 366–381.

Dewalt K, Dewalt B (2011). Participant Observation: A Guide for Fieldworkers, 2nd ed., Lanham, MD: AltaMira.

Ebert-May D, Derting TL, Hodder J, Momsen JL, Long TM, Jardeleza SE (2011). What we say is not what we do: effective evaluation of faculty professional development programs. BioScience *61*, 550–558.

Entwistle A, Entwistle N (1992). Experiences of understanding in revising for degree examinations. Learn Instruct *2*, 1–22.

Freeman S, Haak D, Wenderoth MP (2011). Increased course structure improves performance in introductory biology. CBE Life Sci Educ *10*, 175–186.

Freeman S, O'Connor E, Parks JW, Cunningham M, Hurley D, Haak D, Dirks C, Wenderoth MP (2007). Prescribed active learning increases performance in introductory biology. CBE Life Sci Educ *6*, 132–139.

Freeman S, Parks JW (2010). How accurate is peer grading? CBE Life Sci Educ *9*, 482–488.

Glaser BG, Strauss AL (2010). The Discovery of Grounded theory: Strategies for Qualitative Research, New Brunswick, NJ: Aldine Transaction.

Handelsman J, *et al.* (2004). Scientific teaching. Science *304*, 521–522.

Henderson C (2005). The challenges of instructional change under the best of circumstances: a case study of one college physics instructor. Am J Phys *73*, 778–786.

Henderson C, Beach A, Finkelstein N (2011). Facilitating change in undergraduate STEM instructional practices: an analytic review of the literature. J Res Sci Teach *48*, 952–984.

Herreid CF (2007). Start with a Story: The Case Study Method of Teaching College Science, Arlington, VA: National Science Teachers Association Press.

Hoskins SG, Lopatto D, Stevens LM (2011). The C.R.E.A.T.E. approach to primary literature shifts undergraduates' self-assessed ability to read and analyze journal articles, attitudes about science, and epistemological beliefs. CBE Life Sci Educ *10*, 368–378.

Hoste R (1982). What do examination items test? An investigation of construct validity in a biology examination. J Biol Educ *16*, 51–58.

Knecht KT (2001). Assessing cognitive skills of pharmacy students in a biomedical sciences module using a classification of multiple-choice item categories according to Bloom's Taxonomy. Am J Pharm Educ *65*, 324–334.

Kotovsky K, Hayes JR, Simon HA (1985). Why are some problems hard? Evidence from Tower of Hanoi. Cogn Psychol *17*, 248–294.

Laird TFN, Chen D, Kuh GD (2008a). Classroom practices at institutions with higher-than-expected persistence rates: what student engagement data tell us. New Direct Teach Learn *Fall 2008*, 85–99.

Laird TFN, Shoup R, Kuh G, Schwarz MJ (2008b). The effect of discipline on deep approaches to student learning and college outcomes. Res High Educ *49*, 469–494.

Litzinger TA, Lattuca LR, Hadgraft RG, Newstetter WC (2011). Engineering education and the development of expertise. J Eng Educ *100*, 123–150.

Merriam SB (2009). Qualitative Research: A Guide to Design and Implementation, San Francisco, CA: Jossey-Bass.

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

Murphy KR, Davidshofer CO (2005). Psychological Testing: Principles and Applications, 6th ed., Upper Saddle River, NJ: Pearson Prentice Hall.

National Research Council (2003). BIO2010: Transforming Undergraduate Education for Future Research Biologists, Washington, DC: National Academies Press.

Nolen SB, Haladyna T (1990). Personal and environmental influences on students' beliefs about effective study strategies. Contemp Educ Psychol *15*, 116–130.

Patton MQ (2002). Qualitative Research and Evaluation Methods, 3rd ed., Thousand Oaks, CA: Sage.

Paul RW, Elder L, Bartell T (1997). California Teacher Preparation for Instruction in Critical Thinking: Research Findings and Policy Recommendations, Sacramento, CA: Foundation for Critical Thinking.

Reynolds J, Moskovitz C (2008). Calibrated peer review assignments in science courses: are they designed to promote critical thinking and writing skills? J Coll Sci Teach *38*, 60–66.

Stanger-Hall KF (2012). Multiple-choice exams: an obstacle for higher-level thinking in introductory science classes. CBE Life Sci Educ *11*, 294–306.

Stanger-Hall KF, Shockley FW, Wilson RE (2011). Teaching students how to study: a workshop on information processing and self-testing helps students learn. CBE Life Sci Educ *10*, 187–198.

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

Traub RE, MacRury K (1990). Multiple-choice vs. free response in the testing of scholastic achievement. In: Test und tends 8: jahrbuch der pädagogischen diagnostik, ed. K Ingenkamp and RS Jager, Weinheim, Germany: Beltz Verlag, 128–159.

Wyse AE, Viger SG (2011). How item writers understand depth of knowledge. Educ Assess *16*, 185–206.

Zoller U (1993). Are lecture and learning compatible? Maybe for LOCS: unlikely for HOCS (SYM). J Chem Educ *70*, 195–197.