JOURNAL OF MICROBIOLOGY & BIOLOGY EDUCATION
DOI: https://doi.org/10.1128/jmbe.v18i1.1272

Research

Promoting the Multidimensional Character of Scientific Reasoning †

William S. Bradshaw[1]*, Jennifer Nelson[2], Byron J. Adams[3], and John D. Bell[2]
[1]*Department of Microbiology and Molecular Biology, Brigham Young University, Provo, UT 84602,*
[2]*Department of Physiology and Developmental Biology, Brigham Young University, Provo, UT 84602,*
[3]*Department of Biology, Brigham Young University, Provo, UT 84602*

This study reports part of a long-term program to help students improve scientific reasoning using higher-order cognitive tasks set in the discipline of cell biology. This skill was assessed using problems requiring the construction of valid conclusions drawn from authentic research data. We report here efforts to confirm the hypothesis that data interpretation is a complex, multifaceted exercise. Confirmation was obtained using a statistical treatment showing that various such problems rank students differently—each contains a unique set of cognitive challenges. Additional analyses of performance results have allowed us to demonstrate that individuals differ in their capacity to navigate five independent generic elements that constitute successful data interpretation: biological context, connection to course concepts, experimental protocols, data inference, and integration of isolated experimental observations into a coherent model. We offer these aspects of scientific thinking as a "data analysis skills inventory," along with usable sample problems that illustrate each element. Additionally, we show that this kind of reasoning is rigorous in that it is difficult for most novice students, who are unable to intuitively implement strategies for improving these skills. Instructors armed with knowledge of the specific challenges presented by different types of problems can provide specific helpful feedback during formative practice. The use of this instructional model is most likely to require changes in traditional classroom instruction.

INTRODUCTION

In the past several years there has been a concerted national effort to transform biological science education from a mode focused on acquisition of information to one that emphasizes application, inquiry, and development of the ability to reason scientifically (1). This is a response, in part, to the problem of attrition among undergraduate science majors (2, 3) and concern for American international competitiveness (4, 5). There has been a growing consensus among university-level biology instructors that traditional teaching approaches, including classroom practice (6) and assessments (7, 8), fail with a large number of students and are a major contributing factor to these deficits.

One of the remedies that has proven effective at stemming the loss of undergraduates from science fields is to employ research-based teaching methods (6, 9, 10). Specific strategies consistent with this philosophy include the introduction of authentic research experience into course design (11–14), the development of concept inventories (15–17), careful definition of the intellectual nature of course expectations (18, 19), and the introduction of active learning pedagogy into the classroom (6, 20, 21).

In addition to the acquisition of the most fundamental broad concepts that constitute the discipline, the objectives for a biology course aligned with current reform efforts should include the development of the intellectual skills exercised by its practitioners. For example, Light (22) has determined that a characteristic of college faculty who succeed in "making a difference" is that they teach students to think like professionals in the field. The *Vision and Change* report specifically includes data analysis among the "Core Competencies" that biological science should foster (1):

> All students should understand that biology is often analyzed through quantitative approaches. Developing the ability to apply basic quantitative skills to biological problems should be required of all undergraduates, as they will be called on throughout their lives to interpret and act on quantitative data from a variety of sources. (p. 14)

*Corresponding author. Mailing address: Department of Microbiology and Molecular Biology, Brigham Young University, Provo, UT 84602. Tel: 801-225-8437. E-mail: groverclan@q.com.
Received: 14 November 2016, Accepted: 20 December 2016, Published: 21 April 2017.
†Supplemental materials available at http://asmscience.org/jmbe

We subscribe to the principle of "Backward Design" (23). In this model of course planning, the establishment of learning goals (Step 1) is followed by the design of assessments capable of demonstrating that those goals have been met (Step 2): How can we know if our students have acquired the knowledge and skills we have deemed desirable? This sequence may be less intuitive than proceeding directly to the issue of how to teach (Step 3): What pedagogical strategy should we employ in achieving the goals? In our experience, designing effective assessments for higher-order scientific reasoning is difficult, and all too often relegated to the near bottom of our time and effort priorities. The consequence is often to rely on traditional exam questions that may only monitor short-term retention of factual information (24) or to present problems requiring higher-order reasoning without equipping students sufficiently. Thus, perhaps the most justifiable of student complaints, "The exam I met was not the one I expected and prepared for!"

Alberts, a long-time advocate for reform in science education, has highlighted the endemic problem in biology courses: "superficial 'comprehensive coverage' [in an overly information-rich subject] leaving little room for in-depth learning" (25). In addition to the obvious remedy of reducing subject matter coverage, this perspective also calls for a reevaluation of the assessment items most likely to genuinely measure authentic understanding. Thus, Alberts (26) has called for "a high-profile effort to produce quality assessments that measure student learning" in, for example, the ability to "interpret scientific explanations of the natural world" and "to evaluate scientific evidence." Similar arguments have been articulated by Pellegrino (8). We believe that were this aspect of course design (the use of assessments capable of measuring scientific reasoning) to be widely implemented in college biology courses, this would dramatically alter student learning strategies required for scholastic success.

Recent papers have provided valuable examples of improved exam problems that assess higher-order thinking skills (19, 27). They are intended to measure and promote the ability to apply, analyze, and evaluate—scholastic objectives at the top end of Bloom's taxonomy (28), and are usually presented in a selected response format in order to permit ease in grading. The content of these problems is usually focused on conceptual principles of biology such as chromosomes, meiosis, or natural selection. Consider the following example (27), the first in a set assessing applications to a realistic clinical scenario, which independently requires quantitative reasoning from a knowledge of basic biological facts:

> The DNA from a typical human sperm cell weighs approximately 3.3 picograms (a picogram is 10^{-12} g). If all chromosomes weighed approximately the same, how much does a typical chromosome weigh?
>
> a) 0.07 picograms
> *b) 0.14 picograms*
> c) 0.28 picograms
> d) 75.9 picograms
> e) 151.8 picograms

Less frequently, higher-order exam problems are placed in a research setting that simulates actual scientific practice. The rationale for such a design is to give students intellectual experience in those laboratory protocols and the interpretation of resulting data that, in fact, are the basis for the accepted conceptual foundations of, for example, cell biology (29) or environmental science (30). The assessments described in the present study are of this latter kind, analysis of data in the figures and tables generated from research in cell biology. Moreover, they are presented in a format that requires students to write sentences that capture conclusions validated by those experimental results, a more rigorous and likely more accurate measure of learning than selection among multiple options authored by a teacher (7, 31).

The use of both formative and summative assessments of this kind strongly informs modifications in the conduct of the biology classroom. These are summarized well by Knight and Wood (32) as "interactive engagement and cooperative work in place of some lecturing, while retaining course content by demanding greater student responsibility for learning outside of class." Tanner (10) has explored the role of engagement, exploration, explanation, elaboration, and evaluation in classroom practice, and how the order in which these are experienced can affect the effectiveness of learning. Teaching strategies for developing science process skills have been explored by Coil et al. (33). The techniques they employed include requiring writing, visualizing experimental outcomes, collaborative work, oral communication, more effective studying, and metacognition.

Earlier results from research efforts in our own classroom mirror these models. We have found that exam and practice problems that require students to generate written interpretations of the data generated by authentic experimental research facilitate the acquisition of higher-order thinking skills (34, 35). However, this task has proven to be inherently difficult for students, and demanded changes in classroom practice so as to facilitate improvement. Our data demonstrate that a problem-solving emphasis produces performance gains in a large-enrollment setting featuring an in-class workshop format and additional faculty mentoring sessions (34). We have also shown that feedback that is maximally formative and minimally punitive prolongs student motivation to improve by delaying grade decisions (36). There is evidence that these didactic strategies promote positive attitudes toward a course, greater long-term interest in the discipline (37), and improved self-efficacy (38). Finally, we have validated how efforts to "clone the professor" through formative inquiry exchanges result in long-lasting metacognitive and analytical thinking skills (35).

Throughout the several years of these efforts our goal has been to learn the best ways to help students get better at the task of scientific reasoning. We had always envisioned

data analysis as a uniform, unidimensional exercise to be applied in blanket fashion in any biological research setting. In the present work, we explore an alternative hypothesis, that the interpretation of experimental results is, in fact, a multidimensional task, and that helping students improve in it depends, in part, on identifying and selectively attending to its separate elements.

In the present work, we address the following research questions: 1) How rigorous is the data interpretation task as judged by performance data and student perceptions of difficulty? 2) Do various data-interpretation problems rank-order students in the same way? 3) Is there evidence for the existence of definable elements that distinguish one data-interpretation problem from another? 4) Might selective attention to distinct elements in assessments improve student ability in the multi-dimensional process that is authentic scientific reasoning?

MATERIALS AND METHODS

Course description

Biology 360, Cellular Biology, is an upper-level, three-credit hour course required of several academic programs in the College of Life Sciences at Brigham Young University. During the seven-year time period (2001 to 2007) of these studies, it enrolled from 95 to 200 students in each of the fall and winter semesters and about 50 in a summer term. A class typically consisted of 20% juniors and 80% seniors; approximately 35% were women. The classes were remarkably homogeneous with respect to academic background. The same instructor (author WSB) taught all the classes during this study (some were co-taught with JDB), and an identical course content was presented using a detailed Lecture Outline (which included in-class practice problems) provided to every student. Assistance was provided each semester by two or three teaching assistants (undergraduates with a record of superior performance while previously enrolled in the course). Most frequently, a TA served for a two-year period of time. The text used was Alberts, *Molecular Biology of the Cell*, 3rd and 4th editions (39, 40).

The subject matter of the course covered five major themes: 1) proteins and membranes; 2) organelle function; 3) gene regulation; 4) signal transduction; and 5) developmental regulation. The stated aims of the course were to: a) promote capture of these essential principles of the discipline in the face of voluminous surrounding detail, and b) enhance student ability to interpret experimental results, specifically to construct sentences that correctly draw conclusions validated by the data. The first of these was assessed through traditional conceptual problems requiring students to rehearse their understanding in writing. Examples of the data-interpretation problems used to assess the second aim are found in the Results section below and in Appendix 1 (Supplemental Figures S1–S13). Each supplemental figure is accompanied by the rubric used to evaluate student answers (correct conclusions validated by the data).

Classroom pedagogy

Classroom learning activities were based on the "flipped classroom" model, with emphasis on formative assessment. Accordingly, most of the in-class time was spent helping students assess their understanding of fundamental concepts (acquired through a prior reading assignment) through Socratic dialogue, pair-share exercises, drawing simple diagrams of basic concepts, and short problems that were solved individually or in small groups, then followed by discussion. Many of these short problems were focused on data interpretation in the context of the day's biological topic (see below). The instructor and teaching assistants' role during these exercises was to mingle among the students, actively providing assistance, encouragement, and feedback. Both faculty and teaching assistants also conducted voluntary out-of-class tutorial sessions.

Assessment design

Exam problems. The subject of each data-interpretation problem was a fundamental of cell biology that had been addressed in the course. Each began with a brief description of the experimental setting. The methodologies employed were either familiar to the students or fully explained in these prompts. However, they required cognitive transfer, as the data were presented in a novel, previously unseen setting. The uniform task was "State in one sentence each the conclusions justified by the data." Authoring these problems was generally accomplished by identifying a relevant study from the published literature. For example, the item in Figure S1 came from a paper combining the topics of gene regulation and neuroscience (41). When necessary for calibrating the problem to the level of the course, the data presentation was simplified. In some cases, we altered the form of the data from one that was qualitative (e.g., autoradiograms of histological samples) to one that was more quantitative and could be summarized in a single graph (Fig. 1). Occasionally, we took the creative license to include data from fictitious experiments with results that would be expected based on current understanding of the topic (Fig. S4 is an example of this practice). One or more brief paragraphs were constructed to provide the student with information sufficient to identify what had been done experimentally, but not so much that the problem became trivial or converted to a glorified recall item. Problems used for formative practice in the classroom were identical in design to those presented on exams (as reported in Results below), but were shorter (usually only a single table or figure of data) so that they could be undertaken and critiqued in about 15 minutes. Appendix 1, Figures S9, S11, and S13 are examples.

In our earliest efforts, we presented these problems in a multiple-choice format, but we soon moved to require

written responses. They were calibrated to both the level of the subject matter and to the range of student abilities through multiple trials. In one semester, we compared performance between multiple-choice and essay forms of the same problems. This was achieved through a random distribution of test booklets.

Quizzes and homework. Some additional assignments, performed online or in-class, were evaluated for content or given points strictly for participation. Pre/post comparisons for data analysis skills on non-exam problems were obtained through this route. The data from this source, as reported in Figure 3, were obtained sequentially over several years—a two- to three-semester repetition for each problem. We were experimenting with problems of different complexity during this period. By this same means, we obtained student opinions about the course and its requirements, exemplified by the survey data shown in Table 2. The faculty (WSB, JDB) designed survey items following extensive post-exam scrutiny of the observed deficiencies in student responses. Students were presented choices and selected which element of the data interpretation task was most difficult for them.

Data collection

Performance was assessed through four midterm exams and a final exam. Each of the midterm exams consisted of three conceptual problems requiring broad understanding of basic biological mechanisms, and three data-interpretation problems cast in experimental settings. These exams were administered in the university testing center. The design of the final exam was similar, with a more comprehensive set of problems and some flexibility for students to choose which problems to solve. The final was administered in the course classroom.

Responses to the data-interpretation problems were evaluated by instructors and student raters (usually undergraduate teaching assistants). Student raters were instructed in the grading protocol by course faculty members during training sessions that included practice on several actual exam papers. Data sets in these items typically generated three or four conclusions for a total of 15 points per problem.

The data in the Results section below came from two sources: 1) selected midterm and final-exam course-related data analysis problems described above, and 2) separate (non-exam) pre/post comparisons of performance on non-course-related data sets.

Statistical analyses

Using the software program GENOVA (42), we performed a generalizability analysis (43) to test the reliability of our scoring system: quantify the variance components (student, problem, rater, occasion) in scores on various final exam problems. This demonstrated that 86% of the variability in scores on an exam with multiple questions is determined by overall student ability and the relative difficulty of individual problems; only 3% was attributable to rater variability (the three binomial combinations—e.g., Rater by Problem—and residual accounted for the remaining 11 %; 34). Differences in average scores were assessed by one-way analysis of variance (ANOVA) (with a Bonferroni posttest) or *t*-tests, as appropriate, with $\alpha = 0.05$. Error bars in the figures denote the standard error (SE) unless otherwise indicated.

This project was reviewed by Brigham Young University's Institutional Review Board and granted exempt status.

RESULTS

Data interpretation is a challenging task

For many years, in both formative and summative settings, we have utilized problems designed to promote scientific reasoning. An example of one such instrument, a data-interpretation problem ("Secretion," featuring fictionalized data consistent with proven mechanisms in cell biology), is shown in Figure 1. It is presented in a constructed response (essay) format whose uniform task is "Write in one sentence each the conclusion(s) justified by these data." Correct conclusions constituting the scoring rubric for this problem are found in the legend. Students typically find such questions challenging; the average score on this Secretion problem was 55.2 ± 19.4% (mean ± standard deviation [SD], $n = 452$) over three semesters when administered in a set of three items (with S2 and S3) constituting the final exam. Thirteen additional items of this type, with solutions, are included in Appendix 1, Figures S1–S13.

Table 1 shows the results of student performance on the Secretion problem when presented in a constructed-response format compared with a selected-response (multiple-choice) format using the same stem. In one semester (2001), as a final-exam item, by random distribution, 104 students received the essay version and 52 its multiple-choice equivalent. Percent scores varied among the six valid conclusions (Table 1). However, for four of these, the multiple-choice version averaged 29 points higher. This result and similar ones from other problems of the same type (data not shown) demonstrate, as expected, that in comparison with choosing among options authored by instructors, the more desirable skill of generating and articulating valid conclusions is also the more difficult.

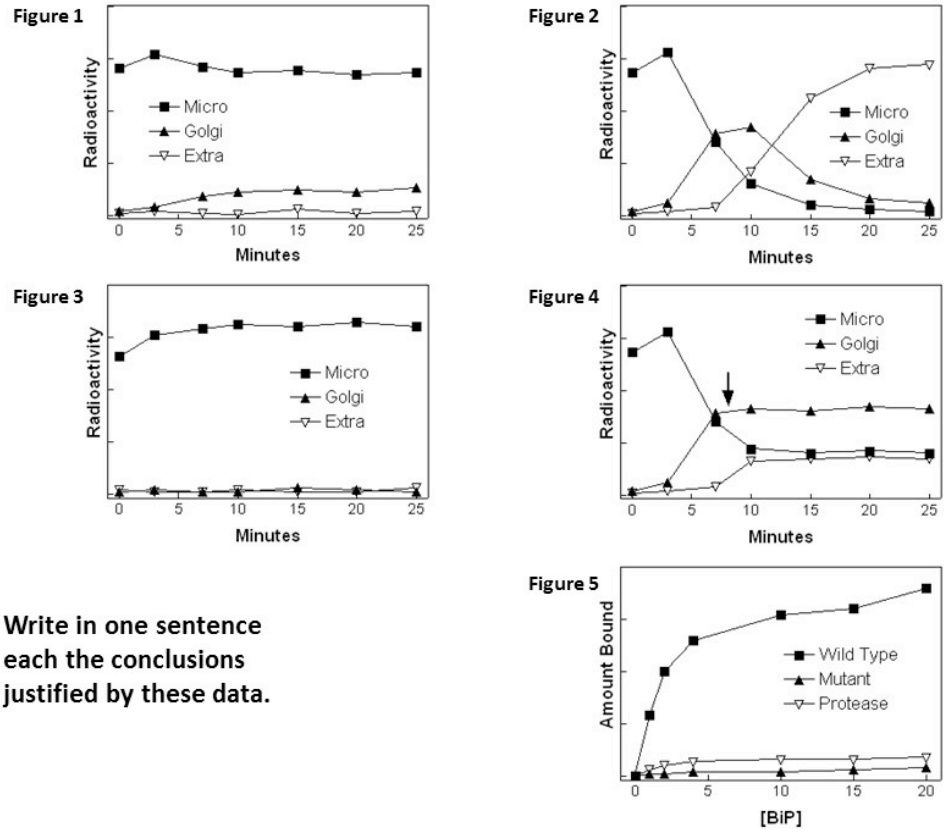
In addition, we solicited student perceptions about which component of the reasoning process is most challenging in a data-interpretation problem. The results, shown in Table 2, are based on written essays explaining each person's choice among the seven options offered (each line of the table represents one of the choices provided in the survey questionnaire). Proceeding intellectually from data to a logical conclusion was cited most often (29%). Reading

comprehension (17%), understanding experimental protocols (15%), and figure and table literacy (16%) were also listed as difficult. The interconnection of several of these elements was cited in many of these responses.

A standard psychometric statistical procedure demonstrated that data-interpretation problems sometimes rank students differently even though the cognitive tasks required by those problems appear superficially to be very similar (same format: brief description of an experimental methods, tables or figures of data, identical task—written conclusion statements). The results of one such generalizability analysis comparing the problem in Figure 1 to two others of the

The chaperone protein BiP was studied in a series of experiments detailed in Figures 1-5 below. In Figures 1-4, cells were pulsed briefly with ^{3}H-leu. At time 0 on the graphs, a chase with non-radioactive leu was initiated. At the time points indicated, cells were frozen, lysed and fractions containing extracellular fluid ("extra"), Golgi or microsomes ("micro") were isolated by differential centrifugation. BiP was then purified from the respective fractions by immunoprecipitation. The data represent the amount of radioactivity in the immunoprecipitates. The data in Figure 1 were obtained with cells expressing wild type BiP. In Figures 2-4, data were obtained with cells that had been engineered to express a mutant form of BiP with amino acid substitutions of K→R and D→N near the C-terminus of the protein. Finally, in Figure 3, the drug brefeldin A (which blocks coatomer-coat assembly) was added prior to time 0. In Figure 4, brefeldin A was added at the time indicated by the arrow.

For Figure 5, microsomes were purified and then disrupted with mild detergent treatment. The binding of purified wild type (squares) or mutant (same cells as for Figures 2-4, triangles) BiP to these microsome membranes was then assessed as indicated in the figure. Lastly, the microsome membranes were treated 30 minutes with a protease. The protease was then inactivated with a specific inhibitor and the binding assay was done using wild type BiP (inverted triangles, "protease").



Write in one sentence each the conclusions justified by these data.

FIGURE 1. Sample data-interpretation problem: Secretion. Valid conclusion statements are: 1) BiP is an ER-resident protein. 2) K and D near the C-terminus are required for return of BiP to the ER from the Golgi. 3) BiP is secreted from the cell if not returned to the ER. 4) Coatomer-coated vesicles are required for the movement of BiP to the Golgi. 5) The default secretory pathway uses coatomer-coated vesicles. 6) BiP binds to a protein in the ER membrane via the K and D sequence.

TABLE 1.
A comparison between student scores on multiple-choice and essay variants of a data-interpretation problem (Secretion).

No.	Secretion Essay Conclusion	EssayMean (%; *n* = 104)	MC Equivalent[a]	MC Mean (%; *n* = 52)	Comments
1	BiP is an ER-resident protein	64.2	i	95.2	identical; true statement
2	C-Term K,D function in return of Bip from Golgi to ER	54.3	h	49.5	false; K,D do not prevent BiP transport to Golgi
			g	61.6	false; not part of signal peptide
3	BiP secreted if not returned from Golgi	42.4	f	79.2	false; not signal for (extracellular) secretion
4	BiP transport to Golgi requires coatomer vesicles	62.3	c	86.3	identical; true statement
5	The export pathway uses coatomer vesicles	52.4	b	45.9	identical; true statement
6	There is a BiP receptor protein in ER membrane	65.2	a	88.5	identical; true statement
			e	61.1	cis to trans transport in Golgi uses coatomer vesicles

[a]Multiple-choice equivalents: the lowercase letters indicate the order in which the choices were presented. Option "e" did not have a counterpart in the Essay version.
MC = multiple choice; ER = endoplasmic reticulum.

TABLE 2.
Self-reported difficulty of the separate steps in solving a data-interpretation problem.

Reasoning Component	No. of Responses	% Responses
Reading comprehension	81	17.0
Application of concepts	33	6.9
Formulation of experimental question	47	9.9
Development of figure literacy	76	16.0
Understanding protocol logic	70	14.7
Reasoning from data to conclusion	139	29.3
Clear written communication	29	6.1

Student responses from fall, 2003, winter 2004, and winter 2005; *n* = 475.

TABLE 3.
Sources of variance in analysis of student scores from the problem in Figure 1 and from multiple problems.[a]

Source of Variation[b]	Percent of Total Variation	
	Secretion Problem	Multiple Problems[c]
Student ability[d]	85.3	21.2
Problem difficulty	N/A	17.9
Student by problem	N/A	44.6
Raters and occasions[e]	5.6	9.0
Residual	3.1	7.3

[a]Analysis performed in fall 2001.
[b]From generalizability analysis.
[c]Items in Figures 1, S1, and S2.
[d]*n* = 156.
[e]Two raters on two occasions.
N/A = not applicable.

same form (shown in Figs. S1 and S2, both utilizing data from the published literature) administered on the same exam identified multiple sources of variance that emphasize these differences (Table 3). For example, 18% of the variation among scores was attributed solely to differences in item difficulty while only 21% was explained entirely by student ability. Interrater reliability was very high. The major factor revealed by the analysis was a student-by-problem interaction (45% of the variation) demonstrating that the three problems ranked students differently. This strongly suggests that each possesses a unique spectrum of sources of difficulty, and that students likewise differ in which of these sources is most challenging.

We conducted analyses to exclude trivial reasons for this student-by-item interaction. For example, the length of time required by students to complete different items was eliminated as a contributing factor. Final exams were three hours in length, and there was no time limit on midterm exams. No correlation between scores and elapsed time for students who spent at least one hour on either type of exam was observed ($r^2 = 0.009$, $p = 0.15$, $n = 223$). Student gender was also not a factor (chi-squared analysis of the interaction by gender, $p = 0.3$, $n = 562$). Finally, errors introduced by raters

or rating occasions were minimal (Table 3; see validation of scoring rubric below).

Elements that distinguish one problem from another

In an effort to identify the specific nature of problem difficulties that segregate students, we closely examined two problems from the second midterm exam that routinely rank students differently (Fig. 2A). The first contained site-directed mutagenesis data relevant to the signal peptide and peptidase cleavage site in the N-terminal sequence of a lysosomal protein (hereafter "targeting," Fig. S3, EX-2). The second problem focused on molecules that modulate cellular migration along the extra-cellular matrix (hereafter "migration," Fig. S4, MMP2). This analysis was prompted when one of us (WSB) was grading these two items in tandem. He noticed that many individuals who performed poorly on the first, did very well on the second. Why? This observation suggested that the differences in student ranking between these problems might lie in the relative difficulty and weight attached to three elements: generic facility in navigating and drawing inferences from figures and tables, understanding the logic of an experimental protocol, and the ability to connect pre-learned biological concepts with the data that validate them. The results of a second grading, designed to test this hypothesis using a rubric based on these three criteria are shown in Figures 2B and C. As the experimental protocol for "targeting" was generic and self-contained in the tabular presentation of the data, the relatively high scores for the first two elements were combined in the analysis of that item (Fig. 2B).

As evident in Figure 2B, the challenging task in "targeting" was to interpret the results in terms of course concepts. This same exam included a recall conceptual problem in which students were asked to diagram the mechanisms involved in protein targeting. Interestingly, nearly half of the students who failed to connect the data to the signal peptide and peptidase cleavage site in the "targeting" setting *did* include those elements in their conceptual response. Specifically, among 338 students tested, 9% (32 students) included both the signal peptide and the peptidase in their conclusions, and 84% of those 32 also included both details in the conceptual rendering. Among the 21% that did not draw a conclusion about the signal peptide, 46% still identified it in their conceptual answer. Of the 70% who did not draw a conclusion about the peptidase, 48% still included it in their conceptual response. Thus, for many students, failure to address the data completely in drawing their conclusions was a lack of connecting the data to the concept rather than an inability to remember and articulate the concept.

In contrast, the principal difficulty for "migration" was complexity in the experimental protocol. Other differences from the "targeting" problem were that student ability to navigate the data could be disassociated from their understanding of the protocols, and the necessary biological background was provided in the prompt rather than relying

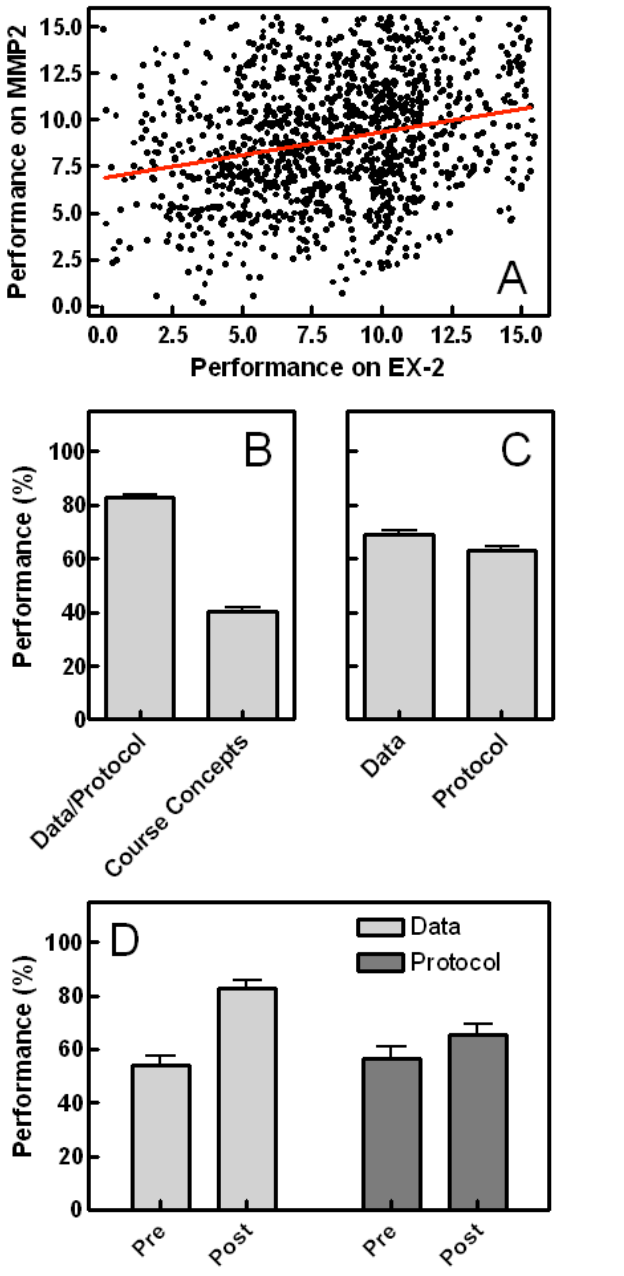


FIGURE 2. Comparison of student performance on two data-interpretation problems. See text for details on the items. (A) Scores on EX-2 and MMP2 were compared over five years (2002–2006) and nine semesters for 1,192 students ($p < 0.0001$ by linear regression, $r^2 = 0.066$). (B) Analysis over consecutive winter and fall semesters (2005) of scores on student responses to EX-2 divided between understanding of data presentation and experimental protocol from those that relied on biological concepts from the course (*t*-test of protocol vs. concepts, $p < 0.0001$, $n = 295$). (C) Responses over the same time period as in (B) to MMP2 were rescored using a rubric that distinguished student understanding of the data presentation from understanding of the experimental protocol (*t*-test of data vs. protocol, $p < 0.01$, $n = 295$). (D) Responses to MMP2 on a test taken at the beginning of a summer term (2006) ("pre") and again during a midterm exam ("post") were scored using the rubric of (C). Two-way analysis of variance: pre vs. post, $p < 0.0001$; data vs. protocol, $p = 0.06$; interaction, $p = 0.01$; $n = 35$.

on recall of conceptual information from the course. This latter assertion was validated by the fact that performance on "migration" was unchanged in a second semester during which the topic of extracellular matrix was intentionally omitted from course lectures (67.8%, topic included; 71.8%, topic not included). Scores for the data navigation and experimental protocol elements are shown in Figure 2C. Both elements were more challenging than the data/protocol element of "targeting" (Fig. 2B). Interestingly, a separate pre/post analysis (in another semester) of performance on "migration" demonstrated a larger course-dependent gain in student ability to apply data navigation skills than in understanding the protocols (Fig. 2D). The former skill thus appears to be teachable over the course of a semester.

Clearly these two data-interpretation problems contain different, definable, sources of difficulty.

In an effort to further resolve these difficulty elements, we created several data-interpretation problems that were designed to contain varied levels of complexity of experimental protocol and dependence on understanding of course concepts. These items were administered to students on a required pretest assignment and then again at the end of the course (actual items shown in Figs. S5–S11). The original intent of some of these was to illustrate principles of learning we wished to make transparent to our students, for example the classical, oft-quoted study in chess players of *pattern recognition* (S5) and the notion of *academic transfer* (S6). These subsequently became part of a set with other items, some of which were nested in a general biology setting, and others whose setting in cell biology was relevant to the course content. Figures 3A and B summarize the results of a pre/post performance comparison, with the problems listed left to right in probable order of increasing number of difficulty elements. The "chess" and "transfer" problems required no more than inference from tabular and graphical data in the context of research scenarios for which the background information was self-contained and unrelated to biology (having played chess or being informed about language learning are not required for a successful analysis). The "sheep" and "prolactin" items also involved deduction from data artifacts but the context was now biological, although not directly dependent on the subject matter of the course. The other three problems ("secretion," "promoter," and "chromatin") required all of the above with the added complexity of connecting specific understanding of biological concepts taught in the course to the data. With respect to experimental protocol, "chess," "transfer," "sheep," "prolactin," and "secretion" all employed simple protocols. "Promoter" and "chromatin" used protocols that are not intuitive without previous exposure and practice (protection assay, DNA footprinting, genetic manipulations).

Significant pre-to-post gains were observed for every item except the "transfer" problem (Fig. 3A). The larger gains for items possessing greater complexity usually reflected bigger decrements in student precourse abilities rather than elevated endpoints. Figure 3B illustrates the fractional gain in score for each item. As emphasized by the horizontal lines, the relative sizes of the pre-to-post gains in scores appeared quantized in synchrony with the number of elements characteristic of each exam problem. The consistency of these quanta among items suggests that the elements they represent are valid. Moreover, this result could explain the observed differences in ranking of students for some problems. For example, students who are challenged by a specific element will struggle with items that emphasize that element but succeed with items that de-emphasize it. Hence, their scores should correlate best for pairs of items that both emphasize or both de-emphasize that element. Accordingly, items that had all four elements correlated with "chromatin" in Figure 3B nine times better than those that had fewer elements ($p = 0.0002$, $n = 4–5$ comparisons). If scores from multiple items relating to a single topic (in this case, gene regulation) are aggregated and compared with those from a different topic (signal transduction) but containing the same number of elements (four), the correlation in student scores became very strong ($r^2 = 0.89$), as displayed in Figure 3C. This is important evidence for the validity of these elements across problems set in different biological subjects; problems of comparable difficulty generate comparable scores.

We now present evidence for an additional diagnostic indicator of assessment elements that differentially affect performance. If such elements exist, and if drawing valid conclusions depends on recognizing and dealing effectively with them, this should be manifest in the shapes of student performance histograms. For most items, these distributions were normal or skewed (see example in Fig. 3D). However, we discovered that occasionally, problems display a very broad or perhaps bimodal distribution as exemplified in Figure 3E. To assess whether the bimodality was simply an artifact of the grading rubric, we invited a faculty expert who had not taught the course nor had access to the scoring rubric to rate student responses to this problem based on his own criteria and understanding of the science. The characteristics of the distribution were retained in this secondary scoring attempt (Fig. 3F, r^2 for correlation between two scoring occasions = 0.87). Close inspection of the problem, which relates to the action of the thyroid hormone receptor (Fig. S12), revealed that all the conclusions students were expected to draw required that they consider all the data in the context of a single unified biological model (in this case, receptor occupies a DNA response element and functions as a repressor until hormone binds and reverses that role). Applying that central conceptual model became a limiting factor in the success of student responses. When students failed to apply the model, they were unable to draw any meaningful conclusions and instead wrote answers that contained disconnected fragments of the truth and offered bizarre explanations without biological precedence. For example, "The thyroid receptor is part of the regulatory [DNA] sequence, because there is still activity with the mutant." These untoward explanations produced

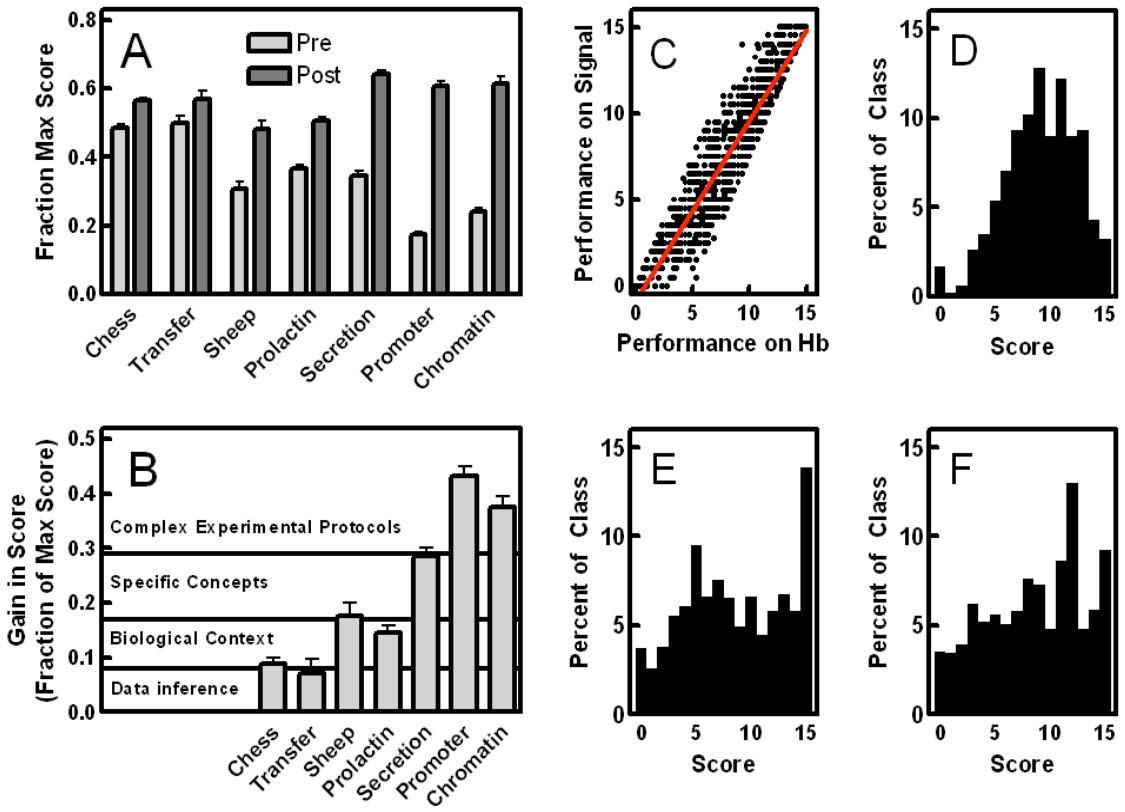


FIGURE 3. (A) Student performance on data interpretation items that vary in complexity (see text for details on the items). Problems were administered at the onset ("Pre") and again near the end of the semester ("Post"), selectively from 2002 to 2006. The main effects (overall pre and post gains, and differences among problems) and the interaction between the two were all significant ($p < 0.0001$ by two-way ANOVA, $n = 111–353$). (B) Gains realized by students for items shown in (A). See text for explanation of labels. (C) Correlation of student performance on two sets of three data interpretation items that contain all four elements from (B). One set focused on hormone signal transduction from a fourth midterm exam, 2006 (signal), and the context for the other was regulation of the globin gene from the third midterm exam (Hb) ($p < 0.0001$ by linear regression, $r^2 = 0.87$). (D–F) Histograms of student performance on an item that displays a normal distribution (D), bimodal distribution (E), and skewed distribution (F). See text for details on these three items.

the low-score mode in the distribution. In contrast, those who applied the model fully were able to address all of the data coherently, producing the high-score mode. Items requiring this approach tended to have multimodal or very broad score distributions.

A careful examination of individual data-interpretation problems, then, demonstrates that they often contain both common features (in varying degrees) and some distinctive features, all of which increase the difficulty of higher-order scientific reasoning (these five elements are highlighted in italics below). All are set in a particular *biological context* with a certain inherent degree of complexity. Each also requires skill in *data inference*. Problem S4 (migration), for example, helps to develop such figure literacy even when one is unfamiliar with the conceptual background in which it is nested. Problem S13 (transport) promotes quantitative thinking, a particularly important aspect of data inference. The data in this assessment illustrate passive transport through a membrane, a concept nearly all of our students could define correctly if asked to do so. Many, however focus only on the shape of the curves, and fail to interpret the meaning of the trials in terms of the *y* axis numbers. They therefore do not make the operational connection to the nature of the process taking place. The experimental results in Problem S3 (targeting) are not sophisticated; the challenge is to *connect the data to course concepts*. As illustrated earlier, Problem S12 (thyroid) assesses a student's ability to *construct a coherent model* of a complex regulatory mechanism (thyroid hormone, hormone receptor, regulatory DNA sequence). Success with Problem S11 (chromatin) requires understanding of *experimental protocols*, an important generic skill used in determining how an answer to a scientific question can be obtained.

DISCUSSION

The results reported here confirm the desirability of using exam problems nested in an authentic research setting and requiring data interpretation in writing to enhance skill in scientific reasoning. That these kinds of assessments are challenging is verified both by performance data, including a comparison with identical stems presented

in a multiple-choice format, and narratives from student surveys. Moreover, a generalizability analysis shows that various problems in a set fail to rank-order students in the same way. Trivial explanations (time on task, gender, grading rubric, or rater differences) fail to explain this outcome. A search for relationships among a set of these problems, however, suggests the existence of separable components of varying degrees of difficulty for various people. The identity of these elements has been made possible through pair-wise analyses of problems in the same exam, differences in the qualitative patterns of score distributions between problems, and performance differences among problems intentionally created to contain different combinations of these putative elements.

Scientific reasoning is difficult for the majority of undergraduate students. It is an analytical skill, not intuitive by nature, that professional biologists acquire through research experience during graduate education. In that setting, the nascent scientist first designs and performs experiments, then determines and defends the meaning of what results, but novice students can benefit vicariously outside the laboratory through practice in the second half of that process. Science is not only multiphasic, but as we have demonstrated, data interpretation is multidimensional intellectually, consisting of separate, if related, elements to be mastered. Acquiring this set of skills has great value, even for those who will never pursue research careers. This becomes self-evident for our students at the same time as they find the training daunting. We frequently hear the following comment, "I really appreciate the improvement in scientific reasoning that the course has afforded me, but I wish I'd been exposed to it earlier and regularly in my scholastic experience." With maturity comes the realization that knowledge acquisition is all too transient, but analytical thinking is much more likely to be persistent and transfer to many aspects of adult life.

Our experience leads us to recommend assessments that require writing. In the context of instructional testing, recognition of the difference between valid and invalid conclusions in a multiple-choice item is positioned at a lower level on a scale of intellectual rigor than the task of independently generating and correctly articulating those interpretations. Thus, "[w]e need to find cost-efficient ways of developing and implementing constructed response tests and performance assessments, as well as a whole array of cognitively sensitive probes of students' understanding to supplement the heavy diet of multiple-choice and short-answer questions so current in today's testing climate" (44). Although greater resources are required to evaluate such items, reliable rating systems are available (34), and the ability to accurately monitor intellectual understanding is considerably greater. Misconceptions are revealed that remain hidden in traditional recall exams.

Psychometric and other analyses of assessments in science courses have identified performance determinants extrinsic to their scientific content, including motivational and situational differences among students (44, 45). The multidimensionality identified here is of a different type. It consists of a unique class of variables, those inherent to the intellectual practice of the discipline itself. If one intent of college-level science courses is to enculturate students into the practice of scientific reasoning, then it makes sense to expose them repeatedly to the diversity of tasks faced by actual practitioners of the discipline. This might be problematic if the sole intent of assessment were reliability in the service of dispensing grades (46, 47). On the other hand such a feature may be a virtue as a reflection of the actual multidimensional experience of practicing scientists. Because these types of problems do not always rank students the same, it is important to include enough of them on exams to generalize overall student performance for the purpose of assigning fair grades. A previous generalizability study indicated that an exam ought to contain at least three of these items. This recommendation is consistent with the high correlation between clusters of three data interpretation items from two separate exams shown in Figure 3C.

Moreover, by having assessment items with diverse emphasis on the elements described herein, and by knowing what that emphasis is, instructors can help students identify specific fundamental weaknesses in their reasoning so that they can make informed efforts to improve. Examples of how this can be accomplished are illustrated in Table 4. Here, the five distinctive dimensions of scientific reasoning which we have identified (constituting a "data analysis skills inventory") are listed, along with the degree to which each is an element in a set of exam problems. The performance on these problems of three hypothetical students is also presented. Consider "Student A." A strong performance on Problems 1, 3, and 4 indicates that this person is generally succeeding in understanding the conceptual principles being taught. However, on Problems 2 and 5, which rely more heavily on sophisticated experimental methods, "Student A" performs poorly. This performance pattern becomes a diagnostic tool that permits the teacher to help this student focus on those specific learning deficits that need to be corrected.

The character of data-interpretation problems used as assessments, both the challenges and the benefits, would seem to demand a reevaluation of classroom practice. Traditional lecturing on selected textbook concepts is unlikely to promote student success with these higher-order tasks. Instead, frequent formative practice with abundant and directed instructor feedback has proven to be highly effective (36, 35). In this mode of instruction, the classroom takes on a workshop-like atmosphere in which students are not permitted to remain passive listeners. The role of peer discussion in this setting is important; in a genetics course, it "enhances understanding, even when none of the students in a discussion group originally knows the correct answer" (48). Similar positive outcomes have been reported for courses in statistics (49) and geoscience (50).

We believe that the benefits of this mode of classroom instruction generalize across a person's scholastic experience

TABLE 4.
Example of how a variety of data interpretation items can be used diagnostically to help students determine strengths and weaknesses in their ability to reason scientifically.

Item	Targeting (S3)	Migration (S4)	Thyroid (S12)	Secretion (Fig. 1)	Chromatin (S11)
Components emphasized					
Data inference	yes	yes	yes	yes	yes
Biological context	yes	yes	yes	yes	yes
Connect to specific concepts from course	yes	no	yes	yes	yes
Complex experimental protocols	no	yes	no	no	yes
Integrate all data with a simple model	not critical	not critical	critical	not critical	not critical
Hypothetical Performance					
Student A	high	low	high	high	low
Student B	low	high	low	low	low
Student C	high	high	low	low	high
Instructor Advice					
Student A	"You need practice in understanding the logic behind experiments."				
Student B	"Draw diagrams of course concepts relevant to the data and use them as a guide for making conclusions."				
Student C	"See if a simple known model can explain all the data before you fabricate new biology."				

and beyond. It places a strong emphasis on a metacognitive appraisal of how one thinks and approaches problem solving (51). Moreover, it provides a model for more effective study outside of class, in which solitary efforts with flash cards and memorization are replaced by interactive group sessions during which individuals are required to articulate their understanding and receive meaningful peer feedback. Students provided with specific direction for both their in-class and out-of-class study efforts no longer have to depend on the ineffective "Well, I guess I'll just have to try harder" prescription. Instead, repeated practice with formative assessments offers a realistic avenue for improvement in the challenging task of scientific thinking.

SUPPLEMENTAL MATERIALS

Appendix 1: Data-interpretation problems used to enhance student ability to interpret experimental results

ACKNOWLEDGMENTS

The contents of this article were developed, in part, under a grant from the United States Department of Education (FIPSE), grant number: P116B041238. The authors have no conflicts of interest to declare.

REFERENCES

1. American Association for the Advancement of Science 2011. Vision and Change in Undergraduate Biology Education: A Call to Action: a summary of recommendations made at a national conference organized by the American Association for the Advancement of Science, July 15–17, 2009. Washington, DC.
2. Seymour E, Hewitt NM. 1997. Talking about leaving: why undergraduates leave the sciences. Westview, Boulder, CO.
3. Chen XL, Soldner M. 2013. STEM attrition: college students' paths into and out of STEM fields. Statistical analysis report. US Department of Education. IES National Center for Education Statistics. Nces.ed.gov/pubs2014/2014001rev.pdf.
4. Bao L, Cai T, Koening K, Fang K, Han J, Wang J, Lu Q, Ding L, Cui L, Lluo Y, Wang Y, Li L, Wu N. 2009. Learning and scientific reasoning. Science 323:586–587.
5. Mervis J. 2010. Shanghai students lead global results on PISA. Science 330:146.
6. Wood WB. 2009. Innovations in teaching undergraduate biology and why we need them. Annu Rev Cell Dev Biol 25:93–112.
7. Stanger-Hall KF. 2012. Multiple-choice exams: an obstacle for higher-level thinking in introductory science classes. CBE Life Sci Educ 11:294–306.
8. Pellegrino JW. 2013. Proficiency in science: assessment challenges and opportunities. Science 340:320–323.
9. Handelsman J, Miller S, Pfund C. 2007. Scientific teaching. WH Freeman and Co, New York, NY.
10. Tanner KD. 2010. Order matters: using the 5E model to align teaching with how people learn. CBE Life Sci Educ 9:159–164.
11. Corwin LA, Graham JJ, Dolan EL. 2015. Modeling course-based undergraduate research experiences: an agenda for further research and evaluation. CBE Life Sci Educ 14:1–13.
12. Brownell SE, Hekmat-Scafe DS, Singla V, Seawell PC, Iman JFC, Eddy SL, Steams T, Cyert MS. 2015. A high-enrollment course-based undergraduate research experience improved student conceptions of scientific thinking and ability to interpret data. CBE Life Sci Educ 14:1–13.

13. Russell JE, D'costa AR, Runck C, Barnes PW, Barrera AL, Hurst-Kennedy J, Sudduth EE, Quinlan EL, Schlueter M. 2014. Bridging the undergraduate curriculum using an integrated course-embedded undergraduate research. CBE Life Sci Educ 14:1–10.
14. Bangera G, Brownell SE. 2014. Course-based undergraduate research experiences can make scientific research more inclusive. CBE Life Sci Educ 13:602–606
15. Couch BA, Wood WB, Knight JK. 2015. The Biology Capstone Assessment: a concept assessment for upper-division molecular biology students. CBE Life Sci Educ 14:1–11.
16. Deane T, Nomme K, Jeffery E, Pollock C, Birol G. 2014. Development of the biological experimental design concept inventory (BEDCI). CBE Life Sci Educ 13:540–551.
17. Brownell SE, Freeman S, Wenderoth MP, Crowe AJ. 2014. BioCore Guide: a tool for interpreting the core concepts of Vision and Change for biology majors. CBE Life Sci Educ 13:200–211.
18. Allen D, Tanner K. 2007. Putting the horse back in front of the cart: using visions and decisions about high-quality learning experiences to drive course design. CBE Life Sci Educ 6:85–89.
19. Crowe A, Dirks C, Wenderoth MP. 2008. Biology in bloom: implementing Bloom's taxonomy to enhance student learning in biology. CBE Life Sci Educ 7:368–380.
20. Linton DL, Pangle WM, Wyatt KH, Powell KN, Sherwood RE. Identifying key features of effective active learning: the effects of writing and peer discussion. CBE Life Sci Educ 2014;13:469–477.
21. Freeman S, Eddy SL, McDonough M, Smith MK, Okoroator N, Jordt H, Wenderoth MP. 2014. Active learning increases student performance in science, engineering, and mathematics. Proc Natl Acad Sci USA 111:8410–8415.
22. Light RJ. 2001. Making the most of college: students speak their minds. Harvard University press, Cambridge, MA.
23. Wiggins G, McTighe J. 1998. Understanding by design. Association for Supervision and Curriculum Development, Alexandria, VA.
24. Momsen JL, Long TM, Wyse SA, Ebert-May D. 2010. Just the facts? Introductory undergraduate biology courses focus on low-level cognitive skills [Reports-Evaluative]. CBE Life Sci Educ 9:435–440.
25. Alberts B. 2013. Failure of skin-deep learning. Science 338:1263.
26. Alberts B. 2009. Restoring science to science education. Iss Sci Technol 25:77–80.
27. Jensen J, McDaniel M, Woodard S. Kummer T. 2014. Teaching to the test… or testing to teach: exams requiring higher-order thinking skills encourage greater conceptual understanding. Educ Psychol Rev 26:307–329.
28. Anderson GL, Krathwohl DR (ed). 2001. A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives. Allyn & Bacon, Boston, MA.
29. Reeve S, Hammond JW, Bradshaw WS. 2004. Inquiry in the large-enrollment science classroom. Simulating a research investigation. J Coll Sci Teach 34:44–48.
30. Bravo A, Porzecanski A, Sterling E, Bynum N, Cawthorn M, Fernandez DS, Freeman L, Ketcham S, Leslie T, Mull J, Vogler D. 2016. Teaching for higher levels of thinking: developing quantitative and analytical skills in environmental courses. Ecosphere 7(4):e01290.
31. Hogan TP, Murphy G. 2007. Recommendations for preparing and scoring constructed-response items: What the experts say. Appl Meas Educ 20:427–441.
32. Knight JK, Wood WB. 2005. Teaching more by lecturing less. Cell Biol Educ 4:298–310.
33. Coil D, Wenderoth MP, Dirks C. 2010. Teaching the process of science: faculty perceptions and effective methodology. CBE Life Sci Educ 9:524–535.
34. Kitchen E, Bell JD, Reeve S, Sudweeks RR, Bradshaw WS. 2003. Teaching cell biology in the large-enrollment classroom: methods to promote analytical thinking and assessment of their effectiveness. Cell Biol Educ 2:180–194.
35. Nelson J, Robison DF, Bell JD, Bradshaw WS. 2009. Cloning the professor, an alternative to ineffective teaching in a large course. CBE Life Sci Educ 8:252–263.
36. Kitchen E, King SH, Robison DF, Sudweeks RR, Bradshaw WS, Bell JD. 2006. Rethinking exams and letter grades: how much can teachers delegate to students? Cell Biol Educ 6:270–280.
37. Kitchen E, Reeve S, Bell JD, Sudweeks RR, Bradshaw WS. 2007. The development and application of affective assessment in an upper-level bell biology course. J Res Sci Teach 44:1057–1087.
38. Reeve S, Kitchen E, Sudweeks RR, Bell JD, Bradshaw WS. 2011. Development of an instrument for measuring self-efficacy in cell biology. J Appl Meas 12:242–260.
39. Alberts B. 1994. Molecular biology of the cell (3rd ed.). Garland Science, New York, NY.
40. Alberts B. 2002. Molecular biology of the cell (4th ed.). Garland Science, New York, NY.
41. Bartsch D, Casadio A, Karl KA, Serodio P, Kandel ER. 1998. Creb1 encodes a nuclear activator, a repressor, and a cytoplasmic modulator that form a regulatory unit critical for long-term facilitation. Cell 95:211–223.
42. Crick JE, Brennan RL. 1982. GENOVA: a generalized analysis of variance system (computer program and manual). American College Testing Program, Iowa City, IA.
43. Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N. 1972. The dependability of behavioral measurement: theory of generalizability for scores and profiles. John Wiley and Sons, Inc., New York, NY.
44. Roeser RW, Shavelson RJ., Kupermintz H. 2002. The concept of aptitude and multidimensional validity revisited. Educ Assess 8:191–205.
45. Lemons PP, Lemons J D. 2013. Questions for assessing higher-order cognitive skills: it's not just Bloom's. CBE Life Sci Educ. 12:47–58.
46. Dunbar SB, Koretz DM, Hoover HD. 1991. Quality control in the development and use of performance assessment. Appl Meas Educ 4:289–303.
47. Shavelson RJ, Baxter GP, Gao X. 1993. Sampling variability of performance assessment. J Educ Measure 30:215–232.

48. Smith MK, Wood WB, Adams WK, Wieman C, Knight JK, Guild N, Su TT. 2009. Why peer discussion improves student performance on in-class concept questions. Science 323:122–124.
49. Goldstein GS. 2007. Using classroom assessment techniques in an introductory statistics class. Coll Teach 55:77–82.
50. Smith G. 2007. How does student performance on formative assessments relate to learning assessed by exams? J Coll Sci Teach 36:28–34.
51. Tanner KD. 2012. Promoting student metacognition. CBE Life Sci Educ 11:113–120.