Undergraduate Biology Lecture Courses Predominantly Test Facts about Science Rather than Scientific Practices

Crystal Uminski,[†‡] Sara M. Burbach,[†] and Brian A. Couch[†*]
[†]School of Biological Sciences, University of Nebraska–Lincoln; Lincoln, NE, 68588; [‡]Thomas H. Gosnell School of Life Sciences, Rochester Institute of Technology; Rochester, NY, 14623

ABSTRACT

Scientific practices are the skills used to develop scientific knowledge and are essential for careers in science. Despite calls from education and government agencies to cultivate scientific practices, there remains little evidence of how often students are asked to apply them in undergraduate courses. We analyzed exams from biology courses at 100 institutions across the United States and found that only 7% of exam questions addressed a scientific practice and that 32% of biology exams did not test any scientific practices. The low occurrence of scientific practices on exams signals that undergraduate courses may not be integrating foundational scientific skills throughout their curriculum in the manner envisioned by recent national frameworks. Although there were few scientific practices overall, their close association with higher-order cognitive skills suggests that scientific practices represent a primary means to help students develop critical thinking skills and highlights the importance of incorporating a greater degree of scientific practices into undergraduate lecture courses and exams.

INTRODUCTION

To address the demands of increasingly interdisciplinary science fields and solve emerging global challenges, education and government agencies have called for undergraduate science courses to emphasize scientific practices (National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, 2007; American Association for the Advancement of Science [AAAS], 2011; National Academies of Sciences, Engineering, and Medicine [NASEM], 2021, 2022). Scientific practices, such as planning investigations, analyzing data, and evaluating information, represent essential skills for establishing, extending, and refining scientific knowledge (National Research Council [NRC], 2007). A robust research synthesis highlighted the importance of scientific practices by naming them as one of the dimensions in a three-dimensional framework for science education (NRC, 2012). In the decade since its publication (NRC, 2012), this three-dimensional framework has quickly risen to prominence in the field of science education (NGSS Lead States, 2013; Cooper *et al.*, 2015; NASEM, 2021) and currently represents the most downloaded report across all publications from the National Academies (Hicks *et al.*, 2022). The three dimensions in the framework consist of *scientific practices* (i.e., the skills students use to engage in science), *crosscutting concepts* (i.e., interdisciplinary ways of thinking about scientific processes), and *disciplinary core ideas* (i.e., concepts central to each science discipline; Table 1). While previous frameworks have featured elements of scientific practices through their emphasis on inquiry (AAAS, 1993; NRC, 1996), these aspects tended to focus on designing investigations and testing hypotheses. The scientific practices included within the three-dimensional framework present a more complete articulation of inquiry and more fully represent the range of actions scientists take to make sense of phenomena (Schwarz *et al.*, 2017).

Tammy Long, *Monitoring Editor*
Submitted Jan 2, 2024; Revised Mar 13, 2024; Accepted Mar 29, 2024
CBE Life Sci Educ June 1, 2024 23:ar19
DOI:10.1187/cbe.23-12-0244
*Address correspondence to: Brian Couch (bcouch2@unl.edu).

TABLE 1. Three-dimensional framework*

Scientific Practices
1. Asking Questions
2. Developing and Using Models
3. Planning Investigations
4. Analyzing and Interpreting Data
5. Using Mathematics and Computational Thinking
6. Constructing Explanations and Engaging in Argument from Evidence
7. Evaluating Information
Crosscutting Concepts
1. Patterns
2. Cause and Effect: Mechanism and Explanation
3. Scale
4. Proportion and Quantity
5. Systems and System Models
6. Energy and Matter: Flows, Cycles, and Conservation
7. Structure and Function
8. Stability and Change
Biology Core Ideas
1. Evolution
2. Information Flow, Exchange, and Storage
3. Structure and Function
4. Pathways and Transformations of Energy and Matter
5. Systems

*Three-dimensional framework adapted from the Three-Dimensional Learning Assessment Protocol (3D-LAP; Laverty *et al.*, 2016). See *Methods* for additional details.

The three-dimensional framework explicitly stresses that students develop deep understanding of science when their learning integrates the three dimensions, rather than approaching them as separate entities (NRC, 2012). Where more traditional science education may emphasize content knowledge (Momsen *et al.*, 2010, 2013), three-dimensional science education provides students opportunities to use scientific practices to extend and refine their understanding of foundational core ideas and crosscutting concepts beyond what can be memorized. Thus, the scientific practices of the three-dimensional framework address the common instructional goal of improving student "critical thinking" abilities (Yuretich, 2003; Stowe and Cooper, 2017). While definitions of critical thinking vary, researchers agree that it represents an essential part of inquiry and involves interpretation, analysis, evaluation, making inferences, and constructing explanations based on evidence (Facione, 1990).

Within undergraduate science education (Crowe *et al.*, 2008), critical thinking has often been identified through Bloom's Taxonomy (Bloom *et al.*, 1956; Anderson *et al.*, 2001). Although limited in its ability to capture the full spectrum of knowledge types (Blumberg, 2009), Bloom's Taxonomy provides a useful tool for classifying cognitive skills that students use when working through a task. The taxonomy is commonly divided into lower-order skills (*remember, understand, apply*) and higher-order skills (*analyze, evaluate, create*). Science education researchers often equate critical thinking with the higher-order skills (Allen and Tanner, 2002; Bissell and Lemons, 2006; Zheng *et al.*, 2008; Moon *et al.*, 2021), and the higher-order skills have considerable parallels to the scientific practices of the three-dimensional framework (Larsen *et al.*, 2022), with some of the same verbs (e.g., *analyze*, *evaluate*) appearing in both frameworks. While they contain considerable overlap, there has not yet been an empirical comparison of scientific practices and Bloom's Taxonomy at the undergraduate level.

The critical thinking contained in Bloom's higher-order cognitive skills have been historically neglected in undergraduate biology courses (Momsen *et al.*, 2010, 2013). Biology instructors often feel the pressure to cover a vast amount of content in their courses (Wright *et al.*, 2018), and while this "breadth over depth" approach exposes students to the core ideas in the discipline, it often only does so at a superficial level that does not provide opportunities to engage in critical thinking. Biology courses that emphasize a wide array of content knowledge may unintentionally reinforce the perception that biology consists of a collection of facts to be memorized (Momsen *et al.*, 2010) and may limit opportunities for students to apply their knowledge to analyze data, develop models, evaluate arguments, design experiments, and participate in other meaningful applications of higher-order cognitive skills and scientific practices.

Scientific practices are at the forefront of K–12 science education, as evidenced by 49 of the U.S. states currently using the three-dimensional framework as the basis for their statewide science standards (NGSS Lead States, 2013; NASEM, 2021; National Science Teaching Association, 2023). Despite this widespread adoption at K–12 levels, there is little evidence indicating to what degree undergraduate biology courses incorporate the three dimensions, particularly with respect to scientific practices. Achieving a smooth transition from high school to undergraduate coursework may depend on the degree to which instruction maintains continuity in three-dimensional language, terminology, and expectations (Clemmons *et al.*, 2020b). Previous efforts have adapted the three-dimensional framework for undergraduate courses (Laverty *et al.*, 2016; Carmel *et al.*, 2019; Bain *et al.*, 2020), marking an important step for further curriculum development and associated research at the college level.

In light of ongoing national calls, there remains a need to determine the extent to which students in undergraduate courses apply the scientific practices outlined in the three-dimensional framework, particularly within the lower-division courses that serve as gateways–and often gatekeepers–to science degree programs (NASEM, 2016). One way to gauge the frequency of scientific practices in a course is to examine course assessments, such as tests and exams. Instructors in lower-division science courses often rely heavily on exams as the primary summative method to measure student learning (Goubeaud, 2010). Because the content of exams inherently reflects the knowledge and skills that instructors value and intend for students to learn (Scouller, 1998; NRC, 2003), an exam including scientific practices signifies that they represent a prioritized learning outcome. This approach of using assessments to gauge the extent of three-dimensional learning in a course has been applied in previous work (Matz *et al.*, 2018; Stowe *et al.*, 2021); however, these studies were conducted using courses taught at a single institution or were limited to three large-enrollment chemistry courses at research-intensive institutions.

The Three-Dimensional Learning Assessment Protocol (3D-LAP; Laverty *et al.*, 2016) is a tool to evaluate the three-dimensional alignment of exams from introductory undergraduate courses in chemistry, physics, and biology. As this protocol is not yet widely implemented in undergraduate biology education, we provide a brief explanation of the 3D-LAP here, with further explanation in the Methods. The 3D-LAP consists of

criteria statements that can be used as indicators of whether an assessment question has the potential to elicit evidence of student engagement with a scientific practice, crosscutting concept, or core idea (Laverty *et al.*, 2016). The scientific practices criteria focus on sources of evidence that indicate students have engaged in the process of science. In contrast, the 3D-LAP criteria for crosscutting concepts and core ideas focus on knowledge itself, with an emphasis on knowledge with explanatory value across or within disciplines, respectively. The crosscutting concept and core idea criteria characterize the specific thinking students might rely upon or the particular disciplinary concept students may have recalled while completing an assessment. For assessment questions to be three-dimensional, they need to meet the 3D-LAP criteria for at least one scientific practice, crosscutting concept, and core idea.

Our study aims to provide the first large-scale, nationwide portrait of how the three-dimensional framework is incorporated into undergraduate biology courses. We use exams as a window into the skills and knowledge instructors prioritize (NRC, 2003), and we analyze exam alignment to the three-dimensional framework with a particular focus on the incorporation of scientific practices. We also analyze exam alignment to Bloom's Taxonomy given its overlap with the science practices of the three-dimensional framework (Larsen *et al.*, 2022) and its wide use in science education (Allen and Tanner, 2002; Crowe *et al.*, 2008; Momsen *et al.*, 2010, 2013; Semsar and Casagrand, 2017; Arneson and Offerdahl, 2018). Our analysis of course exams addresses two research questions: (1) To what extent do exams align to the three-dimensional framework with particular reference to the scientific practices? (2) What is the relationship between an exam's alignment to the three-dimensional framework and to Bloom's Taxonomy of cognitive skills?

MATERIALS AND METHODS

Survey Development and Administration

We developed an online survey through Qualtrics to collect course artifacts (e.g., an exam document, the associated exam answer key, and a syllabus) along with demographic and institutional information from instructors of undergraduate lower-division biology courses. We define lower-division courses as 100- and 200-level courses and their equivalents. To participate in the survey, instructors had to confirm that they were located at a 2- or 4-year institution of higher education in the United States, were currently teaching or had taught a lecture-based lower-division biology course within the past 3 years, and had administered graded tests or exams in their course. We designed these sampling criteria to encompass a wide variety of lower-division biology courses. We provided participants with $75 USD in compensation for the approximately half-hour of time spent completing the survey. This research was classified as exempt from human-subjects review by the University of Nebraska–Lincoln (protocol 21082).

We distributed the survey between May–August 2021 through listservs for professional societies with a focus on undergraduate biology education, including the Society for the Advancement of Biology Education Research (SABER), Ecological Society of America (ESA) EcoEd, Ecological Research as Education Network (EREN), Quantitative Undergraduate Biology Education and Synthesis (QUBES), and National Association of Biology Teachers (NABT). We received 103 survey responses; because of expected overlap in these email lists, we could neither determine the total number of biology instructors who received a survey invitation nor calculate a corresponding response rate across sources. We also wanted to sample from instructors who may not subscribe to education-related listservs, so we randomly selected United States institutions from the Carnegie Classification of Institutions of Higher Education (Indiana University Center for Postsecondary Research, 2021). We randomly selected five institutions from four institution types (i.e., Associate's, Baccalaureate, Master's, and Doctoral) and distributed the survey to all biology instructors at each institution via the email addresses provided on institution websites. We distributed a total of 384 survey invitations using this direct emailing method and received eight responses.

We collected one summative exam from each instructor from a lecture (i.e., nonlab) course. We recognized that instructors may also have used formative assessments and other summative assessments (e.g., projects, papers, presentations) within their courses or in associated labs (e.g., lab reports). Given the variation in the design, format, and grading of these other assessments and lab courses, we only collected lecture exams for this study. We informed instructors that their uploaded materials would be used for research purposes. We did not provide additional specifications or requirements about the type, format, or topic of the course exam. To aid in our coding, we also asked instructors to upload the answer key associated with their exam. We received answer keys from 104 instructors. Instructors were informed that their exams and answer keys would not be shared and that their data would only be presented in aggregate form.

Data Sources

The final dataset contained responses from 111 instructors at 100 unique institutions across the United States, including broad representation from each institution type and geographic region (Table 2). Instructor demographics are in Supplemental Table 1. While the demographics of our sample do not represent the general population (National Science Foundation & National Center for Science and Engineering Statistics, 2019), they do reflect the demographics of biology faculty in the United States (Meixiong and Golden, 2021).

Our sample included different categories of lower-division courses (Supplemental Table 2), with 89 courses being introductory-level and the remaining courses spanning a variety of lower-division biology topics such as anatomy and physiology, environmental science, and microbiology. There were 95 courses that had an associated lab component. With respect to delivery format, 45 courses were taught in-person, five courses were taught online (and had always been structured as online courses), 33 courses had previously been taught in-person but were moved to an online format because of the COVID-19 pandemic, and 28 courses were taught in a hybrid structure containing both in-person and online components. Class sizes ranged from four to 600 students (M = 83.8 ± 10.6 SEM).

Item Coding

We used existing item-coding protocols to code biology exam items for alignment to the three-dimensional framework (Table 1; Supplemental Table 3). We used the Three-Dimensional Learning Assessment Protocol (3D-LAP; Laverty *et al.*, 2016) to

TABLE 2. Institutional Carnegie classifications and geographic regions

Institution region[a]	Associate's[b]	Baccalaureate	Master's	Doctoral	Total
Northeast	4	4	7	6	21
Midwest and Great Plains	6	10	6	7	29
Pacific Northwest	3	2	0	2	7
Southeast	7	9	4	9	29
Southwest	6	0	2	6	14
Total[c]	26	25	19	30	100

[a]Institution regions are based on the Partnership for Undergraduate Life Sciences Education (PULSE) regional network classifications.
[b]Institutional categories are based on Carnegie classifications (Indiana University Center for Postsecondary Research, 2021).
[c]Compared to the reported distribution of institution types in the United States (36% Associates, 20% Baccalaureate, 25% Master's, and 18% Doctoral; Indiana University Center for Postsecondary Research, 2021), our sample slightly underrepresents Associate's institutions and overrepresents Doctoral institutions; however, this comparison is issued with the caveat that we cannot verify how many U.S. institutions offer lower-division biology courses.

characterize scientific practices and crosscutting concepts. Based on the National Research Council's *A Framework for K-12 Science Education* (NRC, 2012), the 3D-LAP was developed as a tool to identify the potential of assessment items in undergraduate science courses to engage students in three-dimensional learning (Laverty *et al.*, 2016).

The 3D-LAP consists of criteria statements regarding whether a student completing an assessment engaged in a scientific practice, crosscutting concept, or core idea. The 3D-LAP contains between two to four criteria statements for each of the scientific practices, and a question is only considered able to engage students in a scientific practice if it meets every criteria statement for that practice. The crosscutting concepts criteria are brief conceptual descriptions, and a question only needs to align with part of the statement to qualify as meeting this dimension. Such is the case for the crosscutting concept "Structure and Function" for which a question may ask a student to either explain a function based on a structure or explain which structure could lead to a specific function (Laverty *et al.*, 2016). Similar to the BioCore Guide (Brownell *et al.*, 2014), the core idea portion of the 3D-LAP contains a main overarching definition of each core idea and a list of how that definition may appear at different biological scales. To qualify as meeting a core idea, a question only needs to align with either the overarching definition or one bullet in the associated list.

We illustrate how the 3D-LAP can be applied to a biology question in Figure 1. This sample question can engage students in the scientific practice "Developing and Using Models" by asking them to interpret two different models of water molecules. We highlight this question as an example of how scientific reasoning can be incorporated into a multiple-choice item and thus address the fourth criteria statement necessary for meeting the "Using Models" scientific practice. This question not only asks students to select which of the two models better represents the interaction of water molecules, the answer options also provide reasoning statements to justify why the model selected is a better representation. To fully engage in the scientific practice of "Developing and Using Models," students must know which is the better model and be able to articulate why. This sample exam question also shows how the 3D-LAP criteria for crosscutting concepts and core ideas are applied. This question illustrates an instance of overlap in the criteria for the crosscutting concept "Structure and Function" and the criteria for the core idea "Structure and Function." The 3D-LAP development team noted that the associated criteria are similar but not identical (Laverty *et al.*, 2016), so these crosscutting concepts and core ideas can be coded independently but are often coded together. As this example question aligns with the criteria for a scientific practice, crosscutting concept, and core idea, we consider this item to be three-dimensional.

There are a few differences between the practices and crosscutting concepts in the original three-dimensional framework and those presented in the 3D-LAP, such as the omission of engineering-specific practices, the merging of the two scientific practices "Constructing Explanations" and "Engaging in Argument from Evidence" into a single practice in the coding protocol, and the separation of the single crosscutting concept "Scale, Proportion, and Quantity" into two independent crosscutting concepts "Scale" and "Proportion and Quantity" (Laverty *et al.*, 2016). These differences reflect a tailoring to the types of assessment tasks observed in undergraduate science courses, and the rationale for these differences is detailed in the initial 3D-LAP development publication (Laverty *et al.*, 2016).

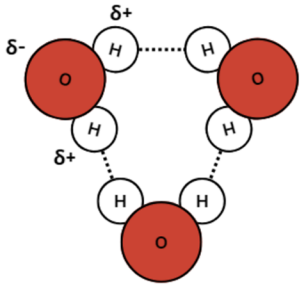
The biology core ideas presented in the 3D-LAP coding protocol only reflect the values and views of disciplinary faculty at one institution, so the 3D-LAP development team encouraged users to substitute other sets of core ideas relevant to their respective disciplines (Laverty *et al.*, 2016). We used the BioCore Guide (Brownell *et al.*, 2014) as the item-coding protocol for the discipline-specific core ideas of the three-dimensional framework. The BioCore Guide delineates the biology core concepts from the *Vision and Change* report across biological scales (AAAS, 2011) and reflects the principles and concepts that a nationwide sample of biology instructors view as important for general biology majors (Brownell *et al.*, 2014).

We used the protocol from Bloom's Dichotomous Key (Semsar and Casagrand, 2017) to assign levels of Bloom's Taxonomy to exam items. We subsequently categorized items with the Bloom's levels *remember*, *understand*, and *apply* as "lower-order cognitive skills" and *analyze*, *evaluate*, and *create* as "higher-order cognitive skills." This binary classification of lower- and higher-order cognitive skills is common in science education research (Zoller, 1993; Fuller, 1997; Bissell and Lemons, 2006; Crowe *et al.*, 2008; Zheng *et al.*, 2008; Freeman *et al.*, 2011).

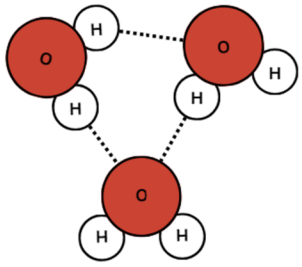
Our sample of 111 exams contained a total of 4337 items (i.e., test questions). Exams ranged from one to 120 items (M = 39.1 ± 2.0 SEM). We used the point values and numbering schemes set by the instructor to determine the boundaries of individual items. Items that shared a common stem, were related to each other as part of a larger task, and/or used a subpart numbering scheme (e.g., 4a, 4b, 4c) were coded as a single clustered item and interpreted as a single unit. This

Two students draw models of water molecules at room temperature. Which student has the model that better explains how water molecules interact at room temperature?

Student 1



Student 2



a) Student 1 because their model better represents hydrogen bonding between hydrogen atoms.
b) Student 1 because their model better indicates the electrostatic differences between oxygen and hydrogen.
c) Student 2 because their model better illustrates covalent bonds between water molecules.
d) Student 2 because their model better shows hydrogen bonding based on the polarity of water molecules.

Dimension	Criteria	Characterization of Item
Scientific Practice: Developing and Using Models	1. Question gives an event, observation, or phenomenon for the student to explain or make a prediction about. 2. Question gives a representation or asks student to select a representation. 3. Question asks student to select an explanation for or prediction about the event, observation, or phenomenon. 4. Question asks student to select the reasoning that links the representation to their explanation or prediction.	1. The question gives the phenomenon of interactions between water molecules at room temperature. 2. The question gives two different representations of water molecules for student to evaluate. 3. The question asks student to select whether Student 1 or Student 2 has the model which better represents the interactions between water molecules at room temperature. 4. The question asks student to select the reasoning linking the representation and the explanation of how water molecules interact.
Crosscutting Concept: Structure and Function	The question asks the student to predict or explain a function or property based on a structure, or to describe what structure could lead to a given function or property.	The question asks student to explain how the polar structure of water leads to properties such as hydrogen bonding between water molecules.
Core Idea: Structure and Function	The three dimensional structure of a molecule and its subcellular localization impact its function, including the ability to catalyze reactions or interact with other molecules.	The question asks student to consider how the structure of water affects its ability to interact with other water molecules.
Bloom's Taxonomy: Analyze	Students are asked to compare/contrast information, have to interpret data (graph, table, figure, story problem, etc.) and come to a conclusion about the data mean, and/or have to decide what data are important to solve the problem (i.e., picking out relevant from irrelevant information).	The question asks student to compare and contrast between two models of water to come to a conclusion about the accuracy of each model in representing water molecules at room temperature.

FIGURE 1. Example question coded for alignment to the three-dimensional framework and Bloom's Taxonomy. This table shows the 3D-LAP (Laverty *et al.*, 2016) criteria for the scientific practice "Developing and Using Models," the 3D-LAP criteria for the crosscutting concept "Structure and Function," and molecular-scale concepts from the BioCore Guide (Brownell *et al.*, 2014) criteria for the core idea "Structure and Function." The table includes criteria modified from the Bloom's Dichotomous Key (Semsar and Casagrand, 2017) for the Bloom's Taxonomy level "Analyze." The third column of the table indicates how the authors applied the criteria statements from the published protocols to code this example item. See Supplemental Table 3 for the full codebook and Supplemental Figures 1–5 for additional examples of how the coding protocols were applied to items from undergraduate biology exams.

approach to coding clusters of items is recommended for the 3D-LAP (Laverty *et al.*, 2016) because a cluster of items may better capture the criteria for a dimension compared with the item subparts in isolation. There were 76 clustered items, comprising less than 2% of the total items.

The mental processes a student engages in when responding to assessment items is context-dependent and is affected by previous instruction or experiences in a course. Thus, we coded the implied cognitive processes targeted by an item based on the apparent intent of the item to elicit specific dimensions or cognitive skills. As we did not have insight into the course content or structure, we included information about instructor expectations from answer keys when coding exam items. We note that this use of answer keys when coding items is a deviation from the original 3D-LAP protocol, but the use of the answer keys can be justified because they provide an important source of course context that might be missing from standalone exam documents.

The 3D-LAP delineates each scientific practice as consisting of nested criteria statements describing different levels within the practice. Similar to previous studies (Laverty *et al.*, 2016; Laverty and Caballero, 2018; Matz *et al.*, 2018; Underwood *et al.*, 2018; Carmel *et al.*, 2019; Stowe *et al.*, 2021), we coded an item as eliciting a scientific practice when it satisfied all of the criteria statements for the corresponding constructed-response or selected-response item type. Some items met the majority of criteria for scientific practices but omitted the reasoning component. Because these items did not fully engage students in a scientific practice (Laverty *et al.*, 2016, 2017), they were not coded as eliciting a scientific practice. We coded an item as addressing a crosscutting concept or core idea if the item aligned with any of the criteria statements within the code. Items may have met multiple scientific practices, crosscutting concepts, or core ideas. We coded only the highest Bloom's Taxonomy level that the item was capable of eliciting. We provide examples of how we applied our codebook to biology exam items in the supplemental materials (Supplemental Figures 1–5). Altogether, our coding procedure reflects 91,093 independent coding decisions across 21 unique codes.

Validity and Reliability

We used an iterative approach to establish the validity and reliability of our coding protocols. Two members of the research team (hereafter Rater 1 and Rater 2) reviewed and discussed the published 3D-LAP, BioCore Guide, and Bloom's Dichotomous Key protocols and agreed that these protocols were representative and appropriate for the breadth of scientific practices, crosscutting concepts, biology core ideas, and cognitive skills expected from lower-division biology exams. Raters 1 and 2 then used these protocols to independently code a training set of 48 items that were randomly selected from the entire item pool. We divided items into four sets of 12 items, and Raters 1 and 2 separately coded the items and then met to discuss and come to consensus on any disagreements before coding the next set of items. In line with the methods from the 3D-LAP coding protocol, we considered the two raters to be in agreement when they agreed on whether an item met a scientific practice, crosscutting concept, core idea, or cognitive skill level (Laverty *et al.*, 2016). As previously established in the 3D-LAP coding protocol and used in other 3D-LAP studies (Laverty *et al.*, 2016; Matz *et al.*, 2018; Carmel *et al.*, 2019), we measured interrater reliability using percent agreement between raters, with a 75% minimum agreement threshold. We calculated percent agreement by adding the number of times the two coders agreed the code was present or agreed the code was absent and dividing by the total number of coded items. We did not use Cohen's Kappa as a measure of interrater reliability for two reasons: (1) this statistic is inappropriate for codes that occur at low or high frequencies (Thompson and Walter, 1988; Brennan and Silman, 1992; Sim and Wright, 2005), and (2) it can be misleading when high agreement is the result of deliberate judgment rather than by chance (Brennan and Prediger, 1981). Our level of agreement between Rater 1 and 2 across the four sets of training items (75–100%; Supplemental Table 4) met or exceeded the published cutoff. Given the acceptable level of agreement, Rater 1 independently coded the remaining items in the sample.

We incorporated a third rater (hereafter Rater 3) to obtain a more robust estimate of interrater reliability and assess potential drift in Rater 1's coding. Rater 3 independently coded the same training set as Raters 1 and 2, similarly discussing disagreements with Rater 1 after each set of 12 items and reaching acceptable agreement (81–100%). Rater 3 then independently coded an additional 487 items (~11% of the sample) from 12 exams. The pseudorandom selection process involved randomly selecting four exams that Rater 1 coded in the first, second, and third phase of their coding process. There was acceptable agreement for each code overall (82–100%) and across the three phases (79–100%), indicating that Rater 1 applied codes similarly throughout the coding process (Supplemental Table 4).

Item Normalization

Given that the exams used different point schemes across courses, for some analyses, we calculated a normalized item point value by dividing the individual item point value by the total number of points on the exam and multiplying by 100. For other analyses, we determined the percent of exam points aligned with the three-dimensional framework or Bloom's Taxonomy levels by summing the normalized item point values on each exam that were aligned to a specific dimension or higher-order cognitive skills.

Analyzing Percent of Exam Points Aligned with each Dimension and Higher-Order Cognitive Skills

We used Pearson correlations and a multiple linear regression to determine the relationships between the percent of exam points aligned with each of the three dimensions and higher-order cognitive skills. The full regression model is included as a note in the associated data table. We used Fisher's z transformations to compare the correlation coefficients with respect to each dimension (Supplemental Table 5). We calculated Pearson correlations and the multiple linear regression using the stats package (v 4.2.3; R Core Team, 2023) and calculated Fisher's z transformations using the diffcor package (v 0.7.1; Blötner, 2022) in R statistical software.

Exam Weighting in Course Grade

Exam weighting can reflect an instructor's perception of the importance of this type of summative assessment within their course. Out of the 111 instructors in the sample, 104 (94%) included a grading scheme that revealed the overall weight of exam grades in their course syllabus. For each course, we

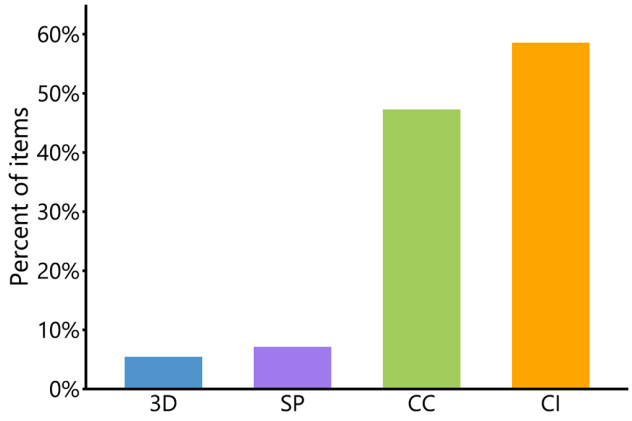


FIGURE 2. Percent of undergraduate biology exam items aligned to each dimension of three-dimensional framework. Exam items ($n = 4337$) are represented only once in each bar even if they may align with multiple scientific practices (SP), crosscutting concepts (CC), or core ideas (CI) within that dimension.

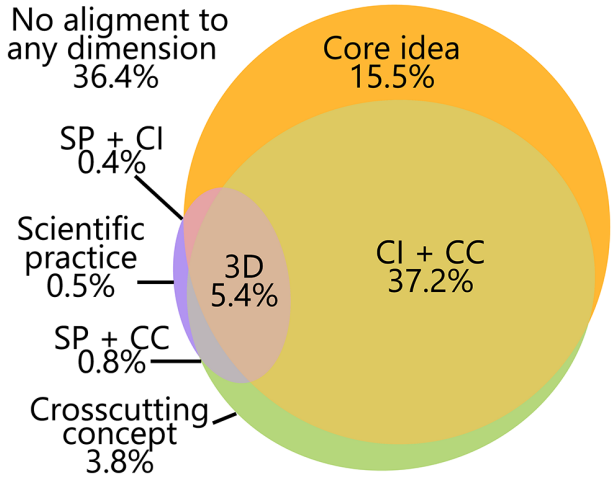


FIGURE 3. Intersections of the three-dimensional alignment of undergraduate biology exam items. The size of the ellipses for scientific practices, crosscutting concepts, and core ideas are proportional to the number of items in the sample aligned with each dimension(s). Approximately 36% of items in the sample did not align with any dimension and are not included within an ellipse.

determined the total percent of the course grade that came from exam grades. We included unit, midterm, and final exams in our value but did not include formative assessments or other summative assessments.

RESULTS

Three-Dimensional Alignment

We collected exams from 111 lower-division biology instructors at 100 unique institutions across the United States (Table 2) and analyzed each exam for three-dimensional alignment. Across our sample of 111 exams with a total of 4337 items (i.e., test questions), only 5% of items ($n = 236$) achieved the principal goal of the three-dimensional framework by simultaneously incorporating a scientific practice, crosscutting concept, and core idea (Figure 2). This lack of three-dimensional alignment was driven by the small percent of items that met the criteria for a scientific practice. Only 7% of items incorporated a scientific practice ($n = 309$), but the majority of those items were three-dimensional (Figure 3). Despite the abundance of items that included a crosscutting concept (47%; $n = 2050$) or core idea (59%; $n = 2540$), only a small proportion of those items qualified as three-dimensional. Strikingly, 36% of items ($n = 1577$) did not align with any of the three dimensions.

When items did align to a scientific practice, the practice was most commonly "Analyzing and Interpreting Data," "Constructing Explanations and Engaging in Argument from Evidence," or to a lesser extent "Developing and Using Models" (Figure 4). While all the scientific practices were represented in the sample, there were notably few items meeting the practices of "Evaluating Information," "Asking Questions," "Planning Investigations," or "Using Mathematics and Computational Thinking." Each crosscutting concept and core idea was represented across the sample. In both the crosscutting concepts and core ideas, "Structure and Function" was the most common code. The codes for "Structure and Function" as a crosscutting concept and as a core idea can be coded independently but given the considerable overlap in the code criteria (Supplemental Table 3), these codes were often applied together.

While the exams contained few items addressing scientific practices overall, these items could have been more involved or taken students more time to complete, thus constituting a larger portion of the exam experience. To address this possibility, we analyzed exam content based on normalized item point values, because instructors tend to assign more points to more substantial items. When accounting for item point value, we found that most exams still had fewer than 10% of points aligned with scientific practices and that 32% of exams had no scientific practices at all (Figure 5). Thus, items targeting scientific practices had higher point values than other exam items, but scientific practices still represented a relatively small proportion of the overall exam content.

Alignment to Bloom's Taxonomy

We applied Bloom's Taxonomy to see which cognitive skills predominated in undergraduate biology exams. We found that the majority of items (92%; $n = 3973$) aligned with the lower-order skills *remember*, *understand*, or *apply*, with just 8% of items ($n = 364$) aligning to the higher-order skills *analyze*, *evaluate*, or *create* (Figure 6). Even after accounting for the tendency of instructors to place more points on higher-order Bloom's items (Welch's ANOVA, $F(1, 368.5) = 42.3$, $p < 0.001$), we found that, overall, exams tended toward lower-order cognitive skills. On average, instructors had 84 ± 2% SEM exam points aligned to lower-order cognitive skills and 16 ± 2% SEM exam points aligned to higher-order cognitive skills (Figure 5). Approximately 30% of instructors did not have any items aligned to higher-order skills on their exam.

Relationship between Three-Dimensional Alignment and Bloom's Taxonomy

There was a strong correlation between the percent of three-dimensional points on an exam and the percent of points targeting higher-order cognitive skills ($r = 0.75$; Figure 7;

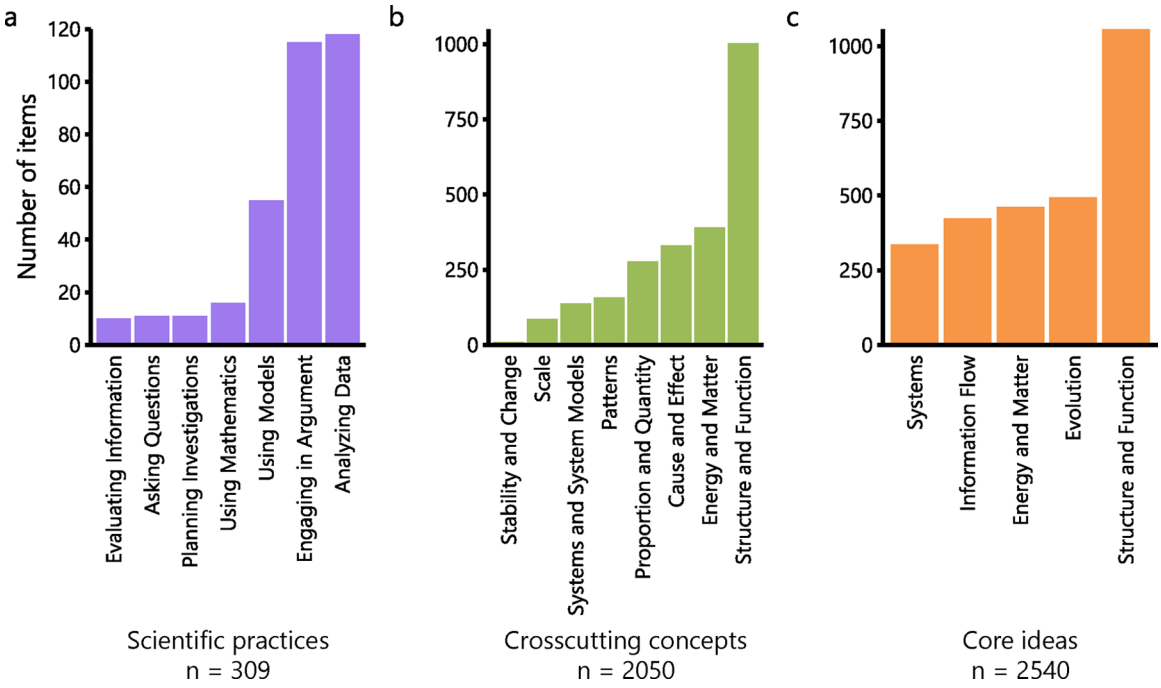


FIGURE 4. Alignment of undergraduate biology exam items to each of the scientific practices, crosscutting concepts, and core ideas of the three-dimensional framework. Individual items may have addressed more than one scientific practice (a), crosscutting concept (b), or core idea (c), thus the sum of the bars in each plot may exceed the total number of items aligned to the dimension. Note the difference in y-axis values for scientific practices relative to crosscutting concepts and core ideas.

Supplemental Table 5). This relationship was driven by scientific practices, which had the highest correlation with Bloom's Taxonomy of any of the three dimensions ($r = 0.88$). Crosscutting concepts and core ideas were also correlated with the percent of higher-order cognitive skills on an exam, albeit to lesser extents ($r = 0.40$ and 0.34, respectively).

Given the differences in their correlations with higher-order cognitive skills, we conducted a multiple linear regression to better understand the nuanced relationship between the three dimensions and Bloom's Taxonomy (Table 3). We found that

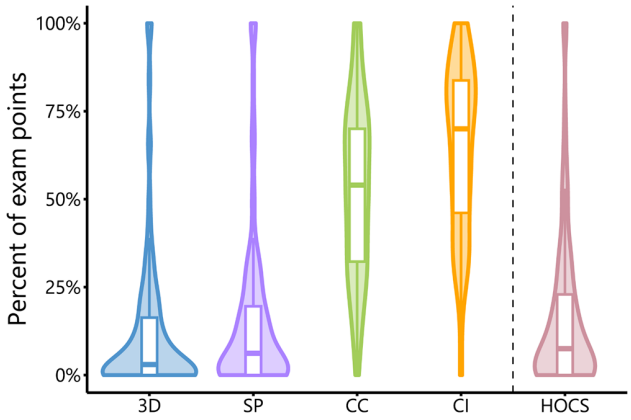


FIGURE 5. Percent of exam points aligned to each dimension of the three-dimensional framework and Bloom's higher-order cognitive skills. For each exam ($n = 111$), we summed the total number of points from items that were aligned to all three dimensions (3D), scientific practices (SP), crosscutting concepts (CC), core ideas (CI), or higher-order cognitive skills (HOCS). Violins represent the distribution of the exam points aligned to each dimension. Solid bars within each box represent the median value, boxes represent the interquartile range, and whiskers represent 1.5 times the interquartile range. An exam is represented once within each dimension.

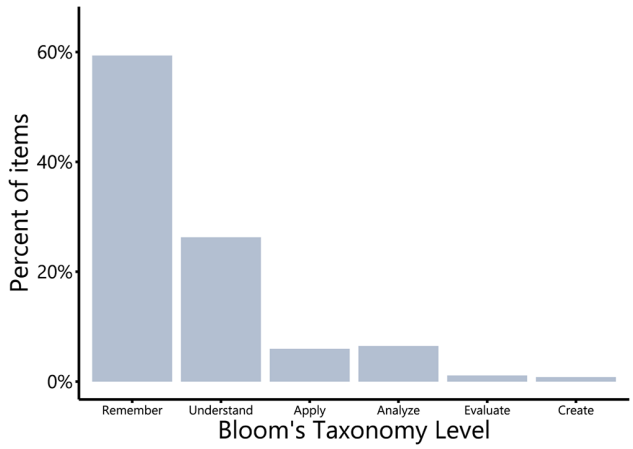


FIGURE 6. Percent of undergraduate biology exam items aligned to each level of Bloom's Taxonomy. Exam items ($n = 4337$) were coded based on the highest level of Bloom's Taxonomy that the item could potentially elicit. The *remember*, *understand*, and *apply* levels of Bloom's Taxonomy are classified here as lower-order cognitive skills, and *analyze*, *evaluate*, and *create* are classified as higher-order cognitive skills.

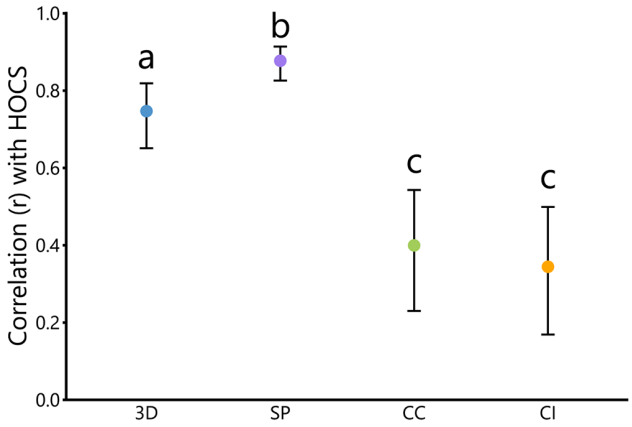


FIGURE 7. Pearson correlation coefficients and 95% confidence intervals representing the relationship between the percent of exam points in each dimension and the percent of exam points assessing higher-order cognitive skills. Letters represent differences in significance between correlation coefficients as determined by Fisher's z-tests (Supplemental Table 5). Abbreviations: 3D = three-dimensional; SP = scientific practice; CC = crosscutting concept; CI = core idea; HOCS = higher-order cognitive skills.

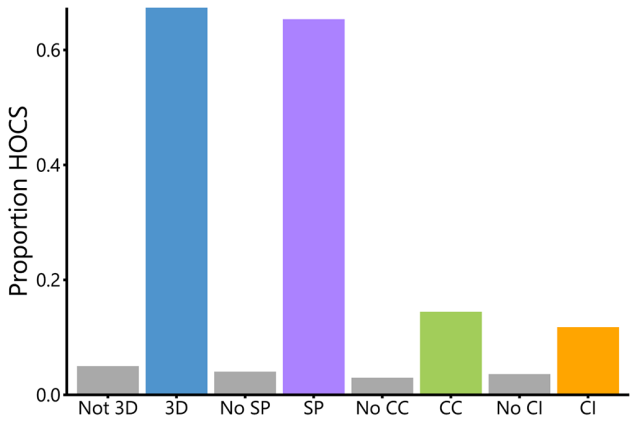


FIGURE 8. Proportion of items in each dimension aligned to higher-order cognitive skills. Colored bars indicate the proportion of items in each dimension that are aligned with the *analyze*, *evaluate*, or *create* higher-order cognitive skills of Bloom's Taxonomy. Proportions are based on the total number of items aligned to all three dimensions ($n = 236$), a scientific practice ($n = 309$), a crosscutting concept ($n = 2050$), and a core idea ($n = 2540$). Gray bars indicate the proportion of items not meeting the given dimension that still aligned with a higher-order cognitive skill. Abbreviations: 3D = three-dimensional; SP = scientific practice; CC = crosscutting concept; CI = core idea; HOCS = higher-order cognitive skills.

only the percent of exam points assessing scientific practices was significantly associated with the percent of points that targeted higher-order skills ($\beta = 0.81$, $t = 16.8$, $p < 0.001$). When considering all three dimensions together, the percent of points assessing crosscutting concepts ($p = 0.92$) and core ideas ($p = 0.54$) did not significantly explain the variance in higher-order skills. These trends are further explained by the proportion of items in each dimension that are aligned with higher-order cognitive skills (Figure 8). The majority of items assessing a scientific practice aligned to higher-order cognitive skills (65%, $n = 202$ of 309 items), whereas only a small percent of items assessing a crosscutting concept or core idea targeted higher-order cognitive skills (14%, $n = 296$ of 2050 items; 11%, $n = 299$ of 2540 items, respectively). Items that did not align to any of the three dimensions rarely targeted the higher levels of Bloom's Taxonomy.

Considering Potential Effects of Additional Course Contexts

Instructors can use other activities to target scientific practices or focus on scientific practices in associated lab courses. However, within our sample, exam grades comprised half of total course grades (M = 49.7% ± 1.5 SEM), and we observed no difference in the extent to which scientific practices (t test, $df = 18.7$, $t = –0.38$, $p = 0.71$) or Bloom's higher-order cognitive skills (t test, $df = 20.3$, $t = –0.38$, $p = 0.70$) were assessed in courses with or without associated labs.

Approximately 65% of the exams ($n = 72$) were administered between March 2020 and the end of our sampling efforts in August 2021, so were affected by the COVID-19 pandemic. While this period of time was marked by changes in instructional modality, with many courses shifting into a partially or fully online format, we did not find notable differences in assessments administered during the global pandemic. When comparing exams administered to students before and after March 2020, we found no significant differences in the percent of three-dimensional points (t test, $df = 65.1$, $t = 1.5$, $p = 0.14$) nor in the percent of higher-order cognitive skills (t test, $df = 68.8$, $t = 0.23$, $p = 0.81$).

DISCUSSION

Taken together, our results highlight a disconnect between what educational reports propose as optimal science assessment (NRC, 2014) and what undergraduate biology lecture courses actually assess on high stakes exams. These reports

TABLE 3. Linear regression model predicting the percent of exam points aligned to higher-order cognitive skills

Effect	Estimate	Standard error	t	p
Intercept	1.21	2.93	0.41	0.68
Points aligned to scientific practices	0.81	0.05	16.83	<0.001
Points aligned to crosscutting concepts	–0.007	0.07	–0.10	0.92
Points aligned to core ideas	0.04	0.06	0.62	0.54

Adjusted $R^2 = 0.76$
Model: Points aligned to higher-order cognitive skills ~ Points aligned to scientific practices + Points aligned to crosscutting concepts + Points aligned to core ideas.

indicate that integrating scientific practices with crosscutting concepts and core ideas is needed for students to reason through how scientific ideas form and to view science as a dynamic and ongoing process (AAAS, 2011; NRC, 2012; NASEM, 2022), but we found that scientific practices are largely missing from exams. The low frequency of science practices paired with the high frequency of items only addressing lower-order cognitive skills means students are more often assessed on conceptual knowledge rather than their ability to apply that information to conduct science. While this study necessarily focused on biology, this phenomenon may be the norm in gateway lecture courses across science disciplines (Stowe and Cooper, 2017; Matz *et al.*, 2018). This exclusion of scientific practices may unintentionally reinforce the perception of science as a collection of discrete facts (NRC, 2012, 2014), which may have negative consequences for retention and persistence of students in science majors (Olson and Riordan, 2012).

Potential Explanations for the Lack of Scientific Practices on Exams

The underrepresentation of scientific practices likely reflects constraints placed on instructors who lack the time, resources, and support for implementing three-dimensional lessons and assessments (NRC, 2014) and who may feel pressured to cover broad ranges of content knowledge (Wright *et al.*, 2018). Another possible explanation for the low frequency is that instructors are incorporating scientific practices in other ways. For example, instructors might be targeting scientific practices through formative assessments (e.g., in-class activities, homework assignments) or other summative assessments (e.g., projects, papers, presentations). Instructors might be preferentially covering scientific practices in associated lab courses. The practices "Evaluating Information," "Asking Questions," "Planning Investigations," and "Using Mathematics and Computational Thinking" occurred least frequently on exams, suggesting that these practices associated with traditional definitions of inquiry may see more prominent implementation in lab courses (Carmel *et al.*, 2019). Conversely, courses without associated labs did not assess more scientific practices, suggesting that the assessment content of the lecture portion of a course may be fairly independent from associated lab sections. Finally, instructors might be reserving instruction and assessment of scientific practices for upper-division courses, yet our previous work found that the extent to which instruction focuses on scientific practices does not differ between lower-division and upper-division courses (Durham *et al.*, 2017). While more research is needed to characterize the extent to which three-dimensional learning occurs in these other places, the presence of three-dimensional components elsewhere does not necessarily negate the importance of scientific practices being incorporated in lecture exams. The three-dimensional framework contends that scientific practices should be incorporated throughout the curriculum because they help students to develop a robust understanding of disciplinary knowledge as the dynamic product of a scientific process.

Most Exam Items Assess Lower-Order Cognitive Skills

Most exam items were only capable of assessing lower-order cognitive skills on Bloom's Taxonomy. The majority of items met the criteria for a core idea or crosscutting concept; however, most of these items did not elicit a scientific practice. Although not true for every case, many of these one- or two-dimensional items tended to ask students to recall definitions or discrete pieces of memorized information (i.e., lower-order cognitive skills). While it is important for students to remember and understand these foundational ideas, the three-dimensional framework calls for students to apply their knowledge and understanding using scientific practices (Cooper *et al.*, 2015). We also observed that over a third of all items did not align with any dimension, signaling that assessments still often contain factual information outside the scope of the biology core concepts. Our work lends credence to the longstanding criticism that lower-division science courses, particularly in biology, overemphasize memorization (Sundberg *et al.*, 1994; Momsen *et al.*, 2010, 2013). Such a finding has consequences for student learning, as memorization-based exams may not be as effective at promoting long-term retention of course content compared with exams that encourage deeper understanding and application (Jensen *et al.*, 2014).

Scientific Practices are a Means to Elicit Higher-Order Cognitive Skills

Many instructors share the goal of teaching and assessing critical thinking and higher-order cognitive skills (Yuretich, 2003), but our findings echo previous studies (Momsen *et al.*, 2010, 2013) and indicate that many instructors may not be meeting that goal. The abundance of lower-order cognitive skills may be in part attributed to a common interpretation of Bloom's Taxonomy in which a high level of item difficulty is conflated with achieving a higher-order Bloom's level (Lemons and Lemons, 2013; Wright *et al.*, 2018; Monrad *et al.*, 2021). The scientific practices offer a way to navigate around this tendency. We found that the extent to which an exam engages students in higher-order cognitive skills associated with critical thinking is closely correlated with the inclusion of scientific practices. This provides support for the idea that incorporating scientific practices represents a more specific way to target the higher-order cognitive skills and associated critical thinking intended by instructors (Stowe and Cooper, 2017).

Three-Dimensional Assessments Provide a Means to Transform Undergraduate Biology Courses

Following principles of Backward Design (Wiggins and McTighe, 2005), assessment content informs instructional design, so striving to incorporate more three-dimensional items into course exams provides a natural impetus for the associated integration of scientific practices into course curricula and instruction. This model of leading course transformation efforts with transformed assessments has been effective for enacting change in gateway science courses (Matz *et al.*, 2018). While we encourage instructors to assess scientific practices on exams, it would often be impractical for exams to consist entirely of three-dimensional items. Three-dimensional assessments require time and effort from instructors to write and grade (Laverty *et al.*, 2016; Furtak, 2017; Nelson *et al.*, 2023), and instructors may value assessing certain content as critical to student advancement within the discipline. Thus, we view any nonzero amount of three-dimensional items as a starting point toward achieving the goals outlined in national calls.

Alternatives to the Three-Dimensional Framework

We applied the three-dimensional framework because of our focus on lower-division courses. The three-dimensional framework is used extensively in K–12 science education and adopting this framework in lower-division courses can help provide a familiar scaffold for students to aid their learning of skills and concepts expected at the undergraduate level. While we use the three-dimensional framework here, other frameworks can be used similarly to characterize important skills and concepts in undergraduate biology courses. The Advanced Placement (AP) Biology Course Framework (College Board, 2020) provides a guide for skills and concepts, but its application may be limited to introductory biology courses. The *Vision and Change* framework (AAAS, 2011) provides a wider lens for program-level learning outcomes that can be applied across all levels of undergraduate biology. Although there are slight differences in terminology, there is substantial overlap between the scientific practices in the three-dimensional framework and the *Vision and Change* core competencies and their articulation within the more delineated BioSkills Guide (Clemmons *et al.*, 2020a,b). For biology courses focused on ecological concepts, instructors may use the 4-Dimensional Ecology Education framework (Berkowitz *et al.*, 2018; Prevost *et al.*, 2019), which in addition to practices, core concepts, and crosscutting themes features another dimension examining human-environment interactions. Each of the aforementioned frameworks can be used to help center curriculum, instruction, and assessments around foundational ideas and skills that are important for scientific literacy, understanding, and participation.

Limitations

While our sample encompassed a broad diversity of instructors and institution types, our results may not be generalizable to all instances of introductory undergraduate biology courses. The majority of our survey participants were recruited through listservs associated with professional societies with a focus on undergraduate biology education, with some additional participants recruited through a direct email approach. Although we could not definitively calculate response rates from listserv recruitment, in both the listserv and direct email distribution methods we estimate that only a small subset of instructors chose to participate. Those who self-selected into this research likely had an interest and/or expertise in biology education, so our sample may not be representative of the assessment practices used across a broader range of undergraduate biology instructors. Given the educational interests of the instructors in our sample, we speculate that our results might overestimate the occurrence of higher-order cognitive skills and three-dimensional alignment in undergraduate biology exams. We note that a similar overestimation was likely present in previous work characterizing the content of undergraduate biology exams (Momsen *et al.*, 2010) as participants in that research had received long-term and intensive training in learner-centered instruction and assessment design before the study. This previous research also found that biology courses primarily assessed recall and comprehension and highlights the need for additional research into how professional development and other forms of instructor training translate into changes to assessment practices.

Our results should also be interpreted in light of our coding procedures. We used the Three-Dimensional Learning Assessment Protocol (3D-LAP; Laverty *et al.*, 2016) as our coding protocol, and the 3D-LAP indicates that for an item to elicit engagement in most scientific practices it must include a reasoning component. As noted previously (Laverty *et al.*, 2017; Carmel *et al.*, 2019) and as we observed in this sample, this reasoning component is often missing from typical assessment tasks. Strictly following the 3D-LAP coding protocol may have obscured when instructors were incorporating elements of scientific practices into their assessments. The presence of these elements of scientific practices, such as incorporating models or mathematical calculations, would not be evident in this coding scheme if the assessments did not include the final reasoning component. In addition, our coding procedure used additional information contained in answer keys. Coding based on the answer keys was a necessary step because it provided insight into instructors' target for student responses that may not have been evident within the item. These implicit expectations were revealed when constructed-response items that did not explicitly state that students needed to explain their thinking had an associated answer key that indicated that the instructor only awarded full credit when responses contained an appropriate explanation or reasoning. In such cases, we considered these items to be able to elicit explanations and reasoning about scientific phenomenon. This method of coding may have potentially overrepresented the degree to which scientific practices were present in constructed-response items, and this bias may have been more pronounced for instructors whose answer keys called for explanation that was not clearly delineated in the question. This discrepancy highlights the importance of instructors adding adequate scaffolding into their constructed-response items so that students provide answers that include the intended reasoning (Hubbard *et al.*, 2017).

Supporting Three-Dimensional Assessment

Our work highlights the need for increased integration of scientific practices and higher-order cognitive skills across science disciplines, in other course components, into lab curricula, and within upper-division courses. A key outcome of science education is to produce students who can think like scientists. To achieve this goal, it is important to critically evaluate what we assess in science courses and to provide instructors with information, resources, and tools that they can use to ensure that their students are engaging with scientific practices.

To facilitate the broader adoption of three-dimensional assessment, we suggest the further development of example items and writing resources appropriate for undergraduate science courses. Instructors wishing to incorporate scientific practices in their exams may find it helpful to consult the Three-Dimensional Learning Assessment Protocol (Laverty *et al.*, 2016). The 3D-LAP provides detailed criteria that can be used to determine whether an exam item has the potential to engage students in scientific practices, and there are guides for using the 3D-LAP to adapt existing exam items (Underwood *et al.*, 2018). Instructors will also need institutional and departmental support, such as providing time and resources for attending professional development and designing new curricular materials (Nelson *et al.*, 2023). The transition into three-dimensional learning and assessment can be challenging and

time-intensive for instructors (Furtak, 2017; Nelson *et al.*, 2023), but it is a task that may lead to more equitable science assessments (Bang *et al.*, 2017; Ralph *et al.*, 2022).

CONCLUSION

The three-dimensional framework represents a major educational advancement because it presents science proficiency as integrating science practices, crosscutting concepts, and core ideas (NRC, 2012). Indeed, scientific knowledge arises from research investigations, so curriculum reform efforts should seek to engage students with conceptual models as evolving products of the science process, rather than invariant truths (Passmore *et al.*, 2009; Zagallo *et al.*, 2016; Matz *et al.*, 2018). Our research suggests that a more direct incorporation of scientific practices represents a key avenue to helping students develop the envisioned integrative proficiency. By focusing on scientific practices within instruction and assessment, we can help cultivate the types of critical thinking needed by scientifically literate citizens and science professionals to tackle global challenges that require both knowledge and action.

ACKNOWLEDGMENTS

We thank participating instructors and Karli Workman for her contributions to project development. Additional thanks to Brittany Demmitt, Meghan Duffy, Cindee Giffen, Lisa Limeri, and Lisa Paciulli for providing research support. This material is based upon work supported by a National Science Foundation (NSF) Graduate Research Fellowship (DGE-1610400) and Improving Undergraduate STEM Education Grant (DUE-2044243). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

Allen, D., & Tanner, K. (2002). Approaches to cell biology teaching: Questions about questions. *CBE—Life Sciences Education*, *1*(3), 63–67. https://doi.org/10.1187/cbe.02-07-0021

American Association for the Advancement of Science. (1993). *Benchmarks for Science Literacy*. New York, NY: Oxford University Press.

American Association for the Advancement of Science. (2011). *Vision and Change in Undergraduate Biology Education: A Call to Action*. Washington, DC: American Association for the Advancement of Science (AAAS). Retrieved from https://live-visionandchange.pantheonsite.io/wp-content/uploads/2011/03/Revised-Vision-and-Change-Final-Report.pdf

Anderson, L. W., Krathwohl, D. R., & Bloom, B. S. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives.* New York, NY: Longman.

Arneson, J. B., & Offerdahl, E. G. (2018). Visual literacy in Bloom: Using Bloom's Taxonomy to support visual learning skills. *CBE—Life Sciences Education*, *17*(1), ar7. https://doi.org/10.1187/cbe.17-08-0178

Bain, K., Bender, L., Bergeron, P., Caballero, M. D., Carmel, J. H., Duffy, E. M., … & Cooper, M. M. (2020). Characterizing college science instruction: The Three-Dimensional Learning Observation Protocol. *PLOS ONE*, *15*(6), e0234640. https://doi.org/10.1371/journal.pone.0234640

Bang, M., Brown, B., Calabrese Barton, A., Rosebery, A., & Warren, B. (2017). Toward more equitable learning in science: Expanding relationships among students, teachers, and science practices. In Schwarz, C. V., Passmore, C., and Reiser, B. J. (Eds.), *Helping Students Make Sense of the World Using Next Generation Science and Engineering Practices* (pp. 33–58). Arlington, VA: National Science Teachers Association Press.

Berkowitz, A. R., Cid, C., Doherty, J., Ebert-May, D., Klemow, K., Middendorf, G., … & Pohlad, B. (2018). *The 4-Dimensional Ecology Education (4DEE) Framework*. Report to the Ecological Society of America. http:/esa.org/4dee

Bissell, A. N., & Lemons, P. P. (2006). A new method for assessing critical thinking in the classroom. *BioScience*, *56*(1), 66–72. https://doi.org/10.1641/0006-3568(2006)056[0066:ANMFAC]2.0.CO;2

Bloom, B. S., Krathwohl, D. R., & Masia, B. B. (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals*. New York, NY: McKay.

Blötner, C. (2022). *diffcor: Fisher's z-tests concerning difference of correlations. Version 0.7.1*.

Blumberg, P. (2009). Maximizing learning through course alignment and experience with different types of knowledge. *Innovative Higher Education*, *34*(2), 93–103. https://doi.org/10.1007/s10755-009-9095-2

Brennan, P., & Silman, A. (1992). Statistical methods for assessing observer variability in clinical measures. *BMJ : British Medical Journal*, *304*(6840), 1491–1494.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient Kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, *41*(3), 687–699. https://doi.org/10.1177/001316448104100307

Brownell, S. E., Freeman, S., Wenderoth, M. P., & Crowe, A. J. (2014). BioCore Guide: A tool for interpreting the core concepts of Vision and Change for biology majors. *CBE—Life Sciences Education*, *13*(2), 200–211. https://doi.org/10.1187/cbe.13-12-0233

Carmel, J. H., Herrington, D. G., Posey, L. A., Ward, J. S., Pollock, A. M., & Cooper, M. M. (2019). Helping students to "do science": Characterizing scientific practices in general chemistry laboratory curricula. *Journal of Chemical Education*, *96*(3), 423–434. https://doi.org/10.1021/acs.jchemed.8b00912

Clemmons, A. W., Timbrook, J., Herron, J. C., & Crowe, A. J. (2020a). BioSkills Guide. *QUBES Educational Resources*. https://doi.org/10.25334/156H-T617

College Board (2020). *P Biology Course and Exam Description, Effective Fall 2020*. https://apcentral.collegeboard.org/media/pdf/ap-biology-course-and-exam-description.pdf

Clemmons, A. W., Timbrook, J., Herron, J. C., & Crowe, A. J. (2020b). BioSkills Guide: Development and national validation of a tool for interpreting the Vision and Change core competencies. *CBE—Life Sciences Education*, *19*(4), ar53. https://doi.org/10.1187/cbe.19-11-0259

Cooper, M. M., Caballero, M. D., Ebert-May, D., Fata-Hartley, C. L., Jardeleza, S. E., Krajcik, J. S., … & Underwood, S. M. (2015). Challenge faculty to transform STEM learning. *Science*, *350*(6258), 281–282. https://doi.org/10.1126/science.aab0933

Crowe, A., Dirks, C., & Wenderoth, M. P. (2008). Biology in Bloom: Implementing Bloom's Taxonomy to enhance student learning in biology. *CBE—Life Sciences Education*, *7*(4), 368–381. https://doi.org/10.1187/cbe.08-05-0024

Durham, M. F., Knight, J. K., & Couch, B. A. (2017). Measurement Instrument for Scientific Teaching (MIST): A tool to measure the frequencies of research-based teaching practices in undergraduate science courses. *CBE—Life Sciences Education*, *16*(4), ar67. https://doi.org/10.1187/cbe.17-02-0033

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction*. Newark, DE: American Philosophical Association.

Freeman, S., Haak, D., & Wenderoth, M. P. (2011). Increased course structure improves performance in introductory biology. *CBE—Life Sciences Education*, *10*(2), 175–186. https://doi.org/10.1187/cbe.10-08-0105

Fuller, D. (1997). Critical thinking in undergraduate athletic training education. *Journal of Athletic Training*, *32*(3), 242–247.

Furtak, E. M. (2017). Confronting dilemmas posed by three-dimensional classroom assessment: Introduction to a virtual issue of Science Education. *Science Education*, *101*(5), 854–867. https://doi.org/10.1002/sce.21283

Goubeaud, K. (2010). How is science learning assessed at the postsecondary level? Assessment and grading practices in college biology, chemistry and physics. *Journal of Science Education and Technology*, *19*(3), 237–245. https://doi.org/10.1007/s10956-009-9196-9

Hicks, D., Zullo, M., Doshi, A., & Asensio, O. I. (2022). Widespread use of National Academies consensus reports by the American public. *Proceedings of the National Academy of Sciences*, *119*(9), e2107760119. https://doi.org/10.1073/pnas.2107760119

Hubbard, J. K., Potts, M. A., & Couch, B. A. (2017). How question types reveal student thinking: An experimental comparison of multiple-true-false and

free-response formats. *CBE—Life Sciences Education*, *16*(2), ar26. https://doi.org/10.1187/cbe.16-12-0339

Indiana University Center for Postsecondary Research. (2021). Bloomington, IN: The Carnegie Classification of Institutions of Higher Education. https://carnegieclassifications.acenet.edu/carnegie-classification/

Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the test…or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review*, *26*(2). 307–329. https://doi.org/10.1007/s10648-013-9248-9

Larsen, T. M., Endo, B. H., Yee, A. T., Do, T., & Lo, S. M. (2022). Probing internal assumptions of the revised Bloom's Taxonomy. *CBE—Life Sciences Education*, *21*(4), ar66. https://doi.org/10.1187/cbe.20-08-0170

Laverty, J. T., & Caballero, M. D. (2018). Analysis of the most common concept inventories in physics: What are we assessing? *Physical Review Physics Education Research*, *14*(1), 010123. https://doi.org/10.1103/PhysRevPhysEducRes.14.010123

Laverty, J. T., Underwood, S. M., Matz, R. L., Posey, L. A., Carmel, J. H., Caballero, M. D., … & Cooper, M. M. (2016). Characterizing college science assessments: The Three-Dimensional Learning Assessment Protocol. *PLoS One*, *11*(9), e0162333. https://doi.org/10.1371/journal.pone.0162333

Laverty, J. T., Underwood, S. M., Matz, R. L., Posey, L. A., Carmel, J. H., Caballero, M. D., … & Cooper, M. M. (2017). Comment on "Analyzing the Role of Science Practices in ACS Exam Items." *Journal of Chemical Education*, *94*(6), 673–674. https://doi.org/10.1021/acs.jchemed.7b00170

Lemons, P. P., & Lemons, J. D. (2013). Questions for assessing higher-order cognitive skills: It's not just Bloom's. *CBE—Life Sciences Education*, *12*(1), 47–58. https://doi.org/10.1187/cbe.12-03-0024

Matz, R. L., Fata-Hartley, C. L., Posey, L. A., Laverty, J. T., Underwood, S. M., Carmel, J. H., … & Cooper, M. M. (2018). Evaluating the extent of a large-scale transformation in gateway science courses. *Science Advances*, *4*(10), eaau0554. https://doi.org/10.1126/sciadv.aau0554

Meixiong, J., & Golden, S. H. (2021). The US biological sciences faculty gap in Asian representation. *The Journal of Clinical Investigation*, *131*(13), e151581. https://doi.org/10.1172/JCI151581

Momsen, J. L., Long, T. M., Wyse, S. A., & Ebert-May, D. (2010). Just the facts? Introductory undergraduate biology courses focus on low-level cognitive skills. *CBE—Life Sciences Education*, *9*(4), 435–440. https://doi.org/10.1187/cbe.10-01-0001

Momsen, J. L., Offerdahl, E., Kryjevskaia, M., Montplaisir, L., Anderson, E., & Grosz, N. (2013). Using assessments to investigate and compare the nature of learning in undergraduate science courses. *CBE—Life Sciences Education*, *12*(2), 239–249. https://doi.org/10.1187/cbe.12-08-0130

Monrad, S. U., Zaidi, N. L. B., Grob, K. L., Kurtz, J. B., Tai, A. W., Hortsch, M., … & Santen, S. A. (2021). What faculty write versus what students see? Perspectives on multiple-choice questions using Bloom's Taxonomy. *Medical Teacher*, *43*(5), 575–582. https://doi.org/10.1080/0142159X.2021.1879376

Moon, S., Jackson, M. A., Doherty, J. H., & Wenderoth, M. P. (2021). Evidence-based teaching practices correlate with increased exam performance in biology. *PLOS ONE*, *16*(11), e0260789. https://doi.org/10.1371/journal.pone.0260789

National Academies of Sciences, Engineering, and Medicine. (2016). *Barriers and Opportunities for 2-Year and 4-Year STEM Degrees: Systemic Change to Support Students' Diverse Pathways*. Washington, DC: National Academies Press https://doi.org/10.17226/21739.

National Academies of Sciences, Engineering, and Medicine. (2021). *Call to Action for Science Education: Building Opportunity for the Future*. Washington, DC: National Academies Press. https://doi.org/10.17226/26152

National Academies of Sciences, Engineering, and Medicine. (2022). *Imagining the Future of Undergraduate STEM Education: Proceedings of a Virtual Symposium*. Washington, DC: National Academies Press. https://doi.org/10.17226/26314

National Research Council. (1996). *National Science Education Standards*. Washington, DC: National Academies Press. https://doi.org/10.17226/4962

National Research Council. (2003). *Assessment in Support of Instruction and Learning: Bridging the Gap Between Large-Scale and Classroom Assessment - Workshop Report*. Washington, DC: National Academies Press. https://doi.org/10.17226/10802

National Academy of Sciences, National Academy of Engineering, and Institute of Medicine. (2007). *Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future*. Washington, DC: National Academies Press. https://doi.org/10.17226/11463

National Research Council. (2007). *Taking Science to School: Learning and Teaching Science in Grades K-8*. Washington, DC: National Academies Press. https://doi.org/10.17226/11625

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: National Academies Press. https://doi.org/10.17226/13165

National Research Council. (2014). *Developing Assessments for the Next Generation Science Standards*. Washington, DC: National Academies Press. https://doi.org/10.17226/18409

National Science Foundation, & National Center for Science and Engineering Statistics. (2019). *Women, Minorities, and Persons with Disabilities in Science and Engineering: 2019* (Special Report NSF 19-304). Retrieved October 25, 2023, from www.nsf.gov/statistics/wmpd

National Science Teaching Association. (2023). *Science Standards*. National Science Teaching Association. Retrieved March 28, 2024, from www.nsta.org/science-standards

Nelson, P. C., Matz, R. L., Bain, K., Fata-Hartley, C. L., & Cooper, M. M. (2023). Characterizing faculty motivation to implement three-dimensional learning. *Disciplinary and Interdisciplinary Science Education Research*, *5*(1). https://doi.org/10.1186/s43031-023-00079-0

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States*. Washington, DC: The National Academies Press. https://doi.org/10.17226/18290

Olson, S., & Riordan, D. G. (2012). Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics. Report to the President. In *Executive Office of the President*. Washington, DC: Executive Office of the President. https://eric.ed.gov/?id=ED541511

Passmore, C., Stewart, J., & Cartier, J. (2009). Model-based inquiry and school science: Creating connections. *School Science and Mathematics*, *109*(7), 394–402. https://doi.org/10.1111/j.1949-8594.2009.tb17870.x

Prevost, L., Sorensen, A. E., Doherty, J. H., Ebert-May, D., & Pohlad, B. (2019). 4DEE—What's next? Designing instruction and assessing student learning. *Bulletin of the Ecological Society of America*, *100*(3), 1–6.

R Core Team. (2023). *R: A language and environment for statistical computing* (4.2.3) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. www.R-project.org/

Ralph, V. R., Scharlott, L. J., Schafer, A. G. L., Deshaye, M. Y., Becker, N. M., & Stowe, R. L. (2022). Advancing equity in STEM: The impact assessment design has on who succeeds in undergraduate introductory chemistry. *JACS Au*, *2*(8), 1869–1880. https://doi.org/10.1021/jacsau.2c00221

Schwarz, C. V., Passmore, C., & Reiser, B. J. (2017). Moving beyond "knowing about" science to making sense of the world. In: *Helping Students Make Sense of the World Using Next Generation Science and Engineering Practices* (pp. 3–21). Arlington, VA: National Science Teachers Association Press.

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education*, *35*(4), 453–472. https://doi.org/10.1023/A:1003196224280

Semsar, K., & Casagrand, J. (2017). Bloom's dichotomous key: A new tool for evaluating the cognitive difficulty of assessments. *Advances in Physiology Education*, *41*(1), 170–177. https://doi.org/10.1152/advan.00101.2016

Sim, J., & Wright, C. C. (2005). The Kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, *85*(3), 257–268. https://doi.org/10.1093/ptj/85.3.257

Stowe, R. L., & Cooper, M. M. (2017). Practicing what we preach: Assessing "critical thinking" in organic chemistry. *Journal of Chemical Education*, *94*(12), 1852–1859. https://doi.org/10.1021/acs.jchemed.7b00335

Stowe, R. L., Scharlott, L. J., Ralph, V. R., Becker, N. M., & Cooper, M. M. (2021). You are what you assess: The case for emphasizing chemistry on chemistry assessments. *Journal of Chemical Education*, *98*, 2490–2495. https://doi.org/10.1021/acs.jchemed.1c00532

Sundberg, M. D., Dini, M. L., & Li, E. (1994). Decreasing course content improves student comprehension of science and attitudes towards science in freshman biology. *Journal of Research in Science Teaching*, *31*(6), 679–693. https://doi.org/10.1002/tea.3660310608

Thompson, W. D., & Walter, S. D. (1988). A reappraisal of the kappa coefficient. *Journal of Clinical Epidemiology*, *41*(10), 949–958. https://doi.org/10.1016/0895-4356(88)90031-5

Underwood, S. M., Posey, L. A., Herrington, D. G., Carmel, J. H., & Cooper, M. M. (2018). Adapting assessment tasks to support three-dimensional learning. *Journal of Chemical Education*, *95*(2), 207–217. https://doi.org/10.1021/acs.jchemed.7b00645

Wiggins, G. P., & McTighe, J. (2005). *Understanding by design, expanded (*2nd ed.). Alexandria VA: Association for Supervision and Curriculum Development.

Wright, C. D., Huang, A., Cooper, K., & Brownell, S. (2018). Exploring differences in decisions about exams among instructors of the same introductory biology course. *International Journal for the Scholarship of Teaching and Learning*, *12*(2). https://doi.org/10.20429/ijsotl.2018.120214

Yuretich, R. F. (2003). Encouraging critical thinking: Measuring skills in large introductory science classes. *Journal of College Science Teaching*, *33*(3), 40–45.

Zagallo, P., Meddleton, S., & Bolger, M. S. (2016). Teaching Real Data Interpretation with Models (TRIM): Analysis of student dialogue in a large-enrollment cell and developmental biology course. *CBE—Life Sciences Education*, *15*(2), ar17. https://doi.org/10.1187/cbe.15-11-0239

Zheng, A. Y., Lawhorn, J. K., Lumley, T., & Freeman, S. (2008). Application of Bloom's Taxonomy debunks the "MCAT myth." *Science*, *319*(5862), 414–415. https://doi.org/10.1126/science.1147852

Zoller, U. (1993). Are lecture and learning compatible? Maybe for LOCS: Unlikely for HOCS. *Journal of Chemical Education*, *70*(3), 195. https://doi.org/10.1021/ed070p195