Exploring Differences In Decisions About Exams Among Instructors Of The Same Introductory Biology Course

Christian D. Wright, Austin L. Huang, Katelyn M. Cooper, and Sara E. Brownell

Arizona State University

(Received 31 October 2017; Accepted 18 January 2018)

College instructors in the United States usually make their own decisions about how to design course exams. Even though summative course exams are well known to be important to student success, we know little about the decision making of instructors when designing course exams. To probe how instructors design exams for introductory biology, we conducted an exploratory interview study with seven instructors teaching the same introductory biology course at a large university. We found that despite designing exams for the same course, instructor exam decisions differed with regard to what content was assessed, the exam format, the cognitive difficulty of exam questions, the resources used when crafting exams, and how exams were administered. We hope that this work can initiate conversations about how college instructors should design exams and lead to more uniformity in how student learning is assessed across the same courses taught by different instructors.

INTRODUCTION

High-stakes course exams are often the dominant way that undergraduates in large-enrollment introductory-level science courses are assessed in the United States (U.S.) (Kendig, 2013). In some courses, students' grades are based exclusively on their performance on summative assessments. Even in courses that have opportunities for low-stakes formative assessment, exams can often make up a significant portion of a student's grade (e.g., Freeman *et al.*, 2007; Smith, 2007).

At large institutions, multiple sections of introductory-level science courses are often taught each term by different instructors (e.g., University of Arizona, 2016; University of California, Los Angeles, 2017). Students sign up for classes with the same course number and the same course description, expecting the same experience. Perhaps more importantly, different offerings of the same course taught by different instructors are viewed as equivalent courses to the degree program, the university, and to future admissions committees for U.S. graduate and professional schools (AAMC, 2018; ADA, 2018). However, instructors often have autonomy over their courses and exams, so students in the same course taught by different instructors may be assessed using markedly different exams.

How different could exams for the same course be? Work out of our research group has shown that the characteristics of exams written by different instructors across multiple sections of the same introductory biology course at one institution were highly variable in the use of open-ended versus closed-ended questions, how cognitively challenging individual questions were, and how difficult questions were, all of which have the potential to influence student learning (Wright *et al.*, 2016). The act of retrieving information during a test can improve student performance on future tests, as well as enhance overall student learning, conceptual organization of knowledge, and transfer of knowledge to novel situations (Roediger *et al.*, 2009, 2011). The types of questions on exams can also influence how students study and how much they learn. Using exams that contain higher cognitive-level questions (e.g., assessed using Bloom's level; Bloom *et al.*, 1956; L. Anderson *et al.*, 2001; Crowe *et al.*, 2008) can help students develop a deeper conceptual understanding of content (Black and Wiliam, 1998; Stanger-Hall, 2012; Jensen *et al.*, 2014). Further, it has been shown that students may learn at the level of their tests rather than the level of instruction (Jensen *et al.*, 2014). When students were taught at a high cognitive level in class, but only assessed on their exams using low cognitive-level questions, they did not develop as much higher-level thinking compared with students who took exams that contained higher cognitive-level questions (Jensen *et al.*, 2014). Exam format can also influence how students study. When students expect open-ended questions, they study content in ways that promote deeper conceptual understanding than when they expect closed-ended questions (Rickards and Friedman, 1978; Thomas and Bain, 1984; Entwistle and Entwistle, 1991). Differences in how exams are designed between two sections of the same course could result in students who take the same course having disparate levels of conceptual understanding at the end of the course.

Additionally, how exams are constructed could differentially affect the performance of different populations of students. We have demonstrated that despite controlling for prior academic ability, women underperformed on introductory biology course exams compared with men as the cognitive difficulty of exams increased (Wright *et al.*, 2016). We also found that as the cognitive difficulty of exams increased, students with low-socioeconomic status underperformed compared with students with middle/high-socioeconomic status (Wright *et al.*, 2016). Students with low-socioeconomic status also underperformed as more open-ended questions were used (Wright *et al.*, 2016). These differences in performance may contribute to observed performance gaps, for example, between male and female students' college science, technology, engineering and mathematics (STEM) exams and course grades (Tai and Sadler, 2001; Kost *et al.*, 2009; Miyake *et al.*, 2010; Rauschenberger and Sweeder, 2010; Creech and Sweeder, 2012; Eddy *et al.*, 2014; Eddy and Brownell, 2016). Thus, instructor decisions in creating exams could differentially influence different groups of students.

Given the degree of college instructor autonomy and that, at large universities, multiple instructors can teach the same course, this can result in variability in exams across different sections of a course, which can lead to inequities in student learning and performance. To our knowledge there are no studies that explore how college instructors are making decisions when constructing

https://doi.org/10.20429/ijsotl.2018.120214

their exams. Due to the potential impact of instructor decisions about exams on students, we conducted an exploratory interview study of instructors who teach different sections of the same introductory biology course at the same institution.

In this research study, we aim to:

1. Explore the decisions that college instructors make when crafting introductory biology exams for the same course

2. Identify rationales that college biology instructors use to support their decisions about exams

METHODS

University and course context

We chose to explore this phenomenon at a large public R1 institution in the southwest United States. We focused our study on instructors teaching the same introductory biology course, BIO101 (this is a pseudonym to protect the identities of the study participants). This institution offers BIO101 at multiple campuses and the course is taught in fall, spring, and summer semesters. This introductory biology course shares the same course number and course description regardless of when or where it is taught. There is neither a common syllabus for the course nor formal coordination among instructors who teach the course. A student's successful completion (e.g., C or higher) of this course is a prerequisite for any upper-level biology course at this institution.

Instructor recruitment

The research team identified a total of 13 different instructors that have taught BIO101 at this institution within the past four years and each of these instructors was sent an individual email inviting them to participate in an interview. Instructors who did not respond within two weeks were sent another email. Our recruitment email asked instructors if they would participate in a 60-minute interview exploring their rationale for why they construct their exams the way they do.

Data collection

Of the 13 instructors who had taught BIO101, 7 instructors agreed to participate in the study (54% response rate). Five instructors taught at one campus and two instructors taught at different branch campuses within the same university. A member of the research team (CDW) conducted semi-structured interviews between May 2016 and August 2016. Interviews lasted between 45 to 60 minutes and were audio recorded and transcribed (Supplemental Table S1). Prior to the interview, the participants completed a short, online survey that asked them to characterize the structure of their BIO101 course as well as a typical exam administered in the most recent iteration of BIO101 that they taught (Supplemental Table S2). The interviewer used the online survey to remind each instructor of their reported exam characteristics (use of low- vs high-order Bloom's questions and open- vs closed-response questions) and instructors were asked to elaborate on their decision-making about these exam characteristics. At the conclusion of the interview, instructors were asked to provide demographic information including their gender identity, their current position/title, whether their position is tenured, tenure-track, or non-tenure-track, the percentage of their work that is attributed to research, teaching, service, and administration, the length of time they have been in their current position, the length of time they have taught biology at the college level, and if they have conducted/published education research. When any responses were unclear or vague, we followed up with instructors at a later date to ensure their responses were accurately interpreted.

This study was done with an approved IRB protocol #00003837.

DATA ANALYSIS

Participant data were de-identified and pseudonyms were given to each participant. Two authors (CDW and ALH) identified themes that emerged from the interviews (Strauss and Corbin, 1990; Kearney, 2001; Glaser and Strauss, 2009). As part of this process, CDW and ALH first constructed a coding rubric together that would be used to identify themes and categories in the interviews. This was an iterative process in which themes and categories in the rubric were molded and transformed with each additional reading of an interview transcript. Once a final rubric was generated (Supplemental Rubric S3), ALH coded each interview using this final rubric. A subset of interviews (10% of the total codes) were coded by CDW in order to establish interrater reliability, reaching a consensus estimate of 85% (Stemler, 2004).

Instructors' quotes have been lightly edited for clarity by inserting clarification brackets or using ellipses to indicate excluded text.

RESULTS AND DISCUSSION

Participant demographics and course characteristics

Instructor pseudonyms, instructor demographic information, and information about each instructor's most recent BIO101 course are presented in **Table 1**.

Instructors teaching BIO101 made different decisions about what to assess on their exams and had different rationales for doing so. We describe key decisions that instructors made below, with quotes from these instructors that illustrate their decisions and why they made them.

How do instructors choose what content to test in BIO101?

Backward design has been widely promoted as a way for instructors to align their assessment with their course goals and activities (Wiggins and McTighe, 1998; Allen and Tanner, 2007; AAAS, 2011; Cooper *et al.*, 2017). The backward design approach suggests that, when designing a course, instructors first outline the learning goals for the course, then identify appropriate assessments to measure the goals and finally, craft a curriculum to achieve those goals (Wiggins and McTighe, 1998). Surprisingly, only a few of the instructors we interviewed *explicitly* stated that they drew from their course learning objectives when constructing their exams. However, instructors who did align their assessments with learning goals focused on broad concepts that they wanted students to know and highlighted that such alignment provided transparency to students about what they should learn, as illustrated by Pete.

> *Pete: "I go back to the learning objectives that I announced for each lecture and I work from there when creating my exam questions [...] Students should know, in general, the kind of mas-*

> *tery we want them to have and almost all the time, it's not listing facts."*

Another instructor, Mia, specifically mentioned that it was important to use learning objectives to design her exams so that exam questions were aligned with her goals for the course; she tells students that every question on her test should align with her learning objectives.

> *Mia: "I tell students that if they ever find a question that doesn't relate to a learning objective, then they can call me on that and I'll drop that question from the test because I think it's really important to tell the students what we are expecting of them and assessing them as they expect."*

Instead of explicitly referencing learning objectives, it was more common for instructors to state that they used their lecture slides when developing exam questions. By using the lecture slides, instructors ensured that what was assessed on exams was covered in lecture. Some instructors were particularly concerned with covering specific subjects rather than covering broader concepts.

> *Hira: "[I use lecture slides] because those are the specific subjects that I talk about and discuss in class. My hope is that [the students] understood what the main subject was and what the important points were. So I'm trying to assess whether they actually understood what I was trying to get through to them."*

Instead of giving students learning goals, Hira asked students to conclude for themselves what is most important from the lecture slides. While lecture slides can be helpful in focusing students on particular topics, it may still be challenging for introductory students to pinpoint the most important topics to study and to distinguish between superficial and deep characteristics of a biology problem presented in class (Smith *et al.*, 2013). Although some instructors used both learning objectives and lecture slides to design questions, the differences among instructors means that the students who are given learning objectives will likely have a much better conception of what an instructor wants them to *know* and be able to *do*, whereas students who can only rely on lecture slides may not know what to expect as far as what they should be able to *do* with the content presented. Consequently, students provided with learning objectives may perform better on assessments than those not provided the same level of transparency. This could result in students with the same level of cognitive mastery of the content ending up with different exam scores in different sections of the same class taught by different instructors.

Instructors also highlighted a common challenge in introductory courses: the depth vs. breadth debate. Because of the ever-expanding nature of what we know in biology, we cannot cover everything in an introductory biology course (AAAS, 2011). Historically, there has been a debate about whether to teach many topics at a more superficial level or a few topics at a much deeper level (Katz and Rath, 1992; R. Anderson, 1995). The instructors that we interviewed had very different approaches to solving this problem despite teaching the same course. Some instructors, such as Pete and Mia, chose to focus on the fundamental concepts or "big ideas" in biology and test those on their exams (AAAS, 2011). For example, Pete talked about testing less on his exams because he has been trying to focus his instruction on diving more in depth into the content to improve student understanding.

> *Pete: "Semester after semester, I think I included fewer topics, trying to go into more depth and to increase understanding rather than recall."*

However, other instructors did not mention focusing on fundamental concepts when making decisions about how to assess students. For example, Geeta talked about feeling the need

Table 1. *Instructor demographics and course characteristics.*

Instructor characteristics							BIO101 characteristics		
Instructor pseudonym	**Gender**	**Position**	**FTE % Breakdown (res-teach-serv-admin / oth)**	**Time teaching as an instructor**	**Semesters teaching BIO101**	**Association with DBER community**	**# of students**	**% of total course points made up by exams**	**TA support for writing /grading exams**
Mia	Female	Associate professor	20-65-15-0%	25 years	9	Yes	450	44%	No
Pete	Male	Associate professor	35-45-20-0%	20 years	3	Yes	350	47%	No
Hira	Female	Instructional professional	0-90-10-0%	4 years	1	No	61	40%	No
Geeta	Female	Lecturer	0-50-0-50%	1.5 years	3	No	395	50%	No
Ted	Male	Senior lecturer	0-100-0-0%	17 years	21	Yes	120	40%	No
Lawrence	Male	Professor	40-40-20-0%	12 years	6	No	456	53%	No
Alex	Male	Associate professor	45-40-15-0%	10 years	3	No	420	50%	No

This table depicts the demographics of instructors teaching BIO101 and the characteristics of their most recent iteration of their BIO101 course. The participants were representative of a variety of professional tenure and non-tenure positions. Participants also had varying levels of responsibilities, teaching experience, and associations with the discipline-based education research (DBER) community. However, most instructors taught in large classrooms with limited TA support. Each instructor was assigned a pseudonym to protect his or her identity.

to cover everything in introductory biology to help prepare students for upper-level courses.

> *Geeta: "It's not fair to focus on one chapter versus the other. All the chapters according to me are important in biology. It's a general introductory biology class. I'm not going to tell them cell biology is more important than population genetics. Everything is important for me in that class and I try to show them that. This is the whole biology range and so I try to tell them that neurobiology is equally important as physiology or anatomy or cell biology or molecular biology. That's why I feel everything has to be given equal weight for all the chapters – the general topics that they should know before they go on to the next 200-level class."*

Despite the national recommendation for biology instructors to present fewer concepts in greater depth (AAAS, 2011), it appears that these instructors held different beliefs about the breadth and depth of the material that should be covered in BIO101. As a result, a student in, for example, Geeta's BIO101 class will be tested on the breadth of her knowledge while a student in Pete's BIO101 class will be tested on the depth of her knowledge, potentially resulting in students who learn different levels of information. Additionally, instructors of upper-level courses may have the challenge of teaching incoming students with very different prior knowledge and skillsets even though students have completed the same pre-requisite course at the same institution.

At what cognitive level do instructors write questions?

We were interested in whether instructors of BIO101 assessed higher versus lower levels of cognitive thinking on their exams, as defined by Bloom's taxonomy (Bloom *et al.*, 1956; L. Anderson *et al.*, 2001; Crowe *et al.*, 2008). Low-level Bloom's questions would include memorization and understanding, while high-level Bloom's questions include application, analysis, evaluation, and synthesis.

There was variability in whether instructors included mostly high-level Bloom's questions or a mix of high-level and low-level Bloom's questions on their exams **(Table 2).** The decision to include higher-order questions was often based on whether higher-level questions were aligned with an instructor's goals of the course, which included getting students to think at a deeper level and/or preparing them for their future careers in biology. When Pete was asked about why his exams contained mostly high-level Bloom's thinking questions, he discussed how his questions aligned with his learning outcomes, which included students learning to apply what they learn in a course.

> *Pete: "Since recall by itself is really low-level learning and [the information] is not likely to stick around for very long, I try to focus on these big concepts, which show up in the learning objectives, which we use as much as possible to design the course. What that means is that repeating back a concept does nothing. Being able to use that concept in an example we didn't talk about is really important. I mean, a student's ability to take that knowledge with them and develop skills so that they have to analyze the questions and learn something, that's valuable for when they leave the class."*

Pete went on to explain how his learning goals were influenced by a national report on biology teaching, *Vision and Change* (AAAS, 2011; Brownell *et al.*, 2014) which in turn influenced how he constructed his exams.

> *Pete: "[I incorporate higher Bloom's-level questions] because I was influenced by Vision and Change, looking at the core concepts, but also looking at the competencies. I do think that the competencies that we're identifying are very important and there's not a whole lot there about memorization. That's really what I'm getting at is the, trying to get at long-term learning and long-term development of skills, analytical skills that students are going to need in the future. "*

Geeta said that her reason for using mostly higher-order questions on her exams was to help students become more competent biologists.

> *Geeta: "I'm somebody who believes that [introductory students] should understand the definition, but they should be able to apply it. It doesn't matter if you know the definition of a gene. They should be able to understand functionally what is the gene doing, not just what is the definition of a gene. They should know what is a gene doing in their body, so that is key. If they are going to be biologists, I feel that they need to know or understand the functionality of some things, not just the definition."*

However, some instructors mentioned that they intentionally chose to include lower-level Bloom's questions on their BIO101 exams because they perceived that lower-level questions help students to develop a foundation of knowledge necessary for them to be biologists. They also felt low-level questions were essential for training and were needed before students could advance to higher levels of Bloom's; this sentiment was illustrated by Ted.

> *Ted: "At this introductory level, a lot of what I think students need to master is simply a mastery of the vocabulary."*

Alex disagreed and positioned that low-level Bloom's questions were not going to help prepare students for their future roles as citizens or scientists.

> *Alex: "One way of testing is to have the students memorize. Obviously, this will not help when they get their degree and go out into society. If they're doing the research that I am, or any field, they have to think, synthesize, analyze."*

However, Alex still chose to include memorization questions on his BIO101 exams. Even though Alex did not think that students needed to know memorization questions for their future careers, he still valued more straightforward memorization questions on his exams as a way of holding students accountable for coming to class.

> *Alex: "Generally, in terms of the difficulty of questions, I try to include some questions that would be really straightforward, something that I repeated in class many times, so students that come in class will definitely get it, it was obvious, it was repeated more than once."*

Some of the instructors who we interviewed frequently conflated difficulty with high cognitive-level questions, assuming that memorization questions would be easier for students to answer than application questions. It is possible to write difficult memorization questions (e.g., asking students to recall obscure facts) and to write easy application questions (e.g., asking students to answer an application question that is very similar to a question they practiced many times in class). Bloom's level does not indicate whether a question is easy or hard, but rather the cognitive level that we expect students to achieve; higher-level Bloom's questions would be more similar to the level of thinking

***Table 2.** Instructor decisions*

Decision	Mia	Pete	Hira	Geeta	Ted	Lawrence	Alex
Content to test							
To reference learning outcomes when deciding what to test	x	x					x
To reference lecture slides when deciding what to test			x		x		x
To limit the number of concepts tested	x	x					
To include fundamental concepts	x	x			x	x	x
Question characteristics							
To include mostly close-ended questions*	x	x	x	x	x	x	x
To include mostly high level Bloom's questions*	x	x		x			
To include a mix of Bloom's level questions*			x		x	x	x
Materials used to construct exams							
To draw from past exam questions	x	x	x	x	x	x	x
To draw from test bank questions	x		x				x
To draw from questions previously presented to students in class			x		x	x	x
To write new questions		x	x	x			x
Exam delivery							
To create in-class, closed-book exams and out-of-class open-book exams as opposed to only in-class, closed-book exams						x	x

Decisions that instructors made while constructing exams. This table was crafted based on instructor responses during the interview and in the online survey and illustrates the variety of decisions that instructors explicitly stated that they made while crafting their exams. There were four major types of decisions instructors indicated they made: decisions about what content to test, decisions about the characteristics of individual questions, decisions about the materials used to construct exams, and decisions about the format and delivery of the assessments. The "*" indicates decisions that instructors were explicitly asked about during the interview and the other decisions emerged from coding the interviews overall.

like a scientist (Brownell *et al.*, 2015). However, instructors like Hira conflate Bloom's level and difficulty when they suggest that including high-level Bloom's questions increase the chance that a student will fail; it has been established that students can find high-level Bloom's questions easy if they have enough practice solving similar problems (Brownell *et al.*, 2015).

> *Hira: "That's my goal that [students] will be able to analyze and synthesize and think critically about the subjects we are talking about, not just reciting the information that I give them. So, that is my overall goal and I wish I could extend that even more but then I don't want most of the class to fail. So, I'm trying to find a balance between [student failing] and really pushing them to think, at least during the exam."*

The theme of wanting to prevent students from being discouraged came up numerous times, with several instructors discussing how their perceptions of students' backgrounds, study habits, and ability to think critically led to their conclusion that incorporating too many higher-order questions would result in students failing the exam and thus discourage them.

> *Ted: "I think that only high-level questions - that level of an exam is a little bit much to ask out of a first-year student. We want to start them thinking, but let's be honest. They have had probably absolutely no training in grade school or high school on critical thinking skills. Even if they were trained at that age, they probably didn't pay attention, so I think it's difficult for them. I certainly get the feedback from students that the kind of questions I ask that ask them to analyze and think are very difficult. I fear that if you make everything really hard, you will discourage too many students."*

> *Hira: "I try to limit the number of high-level questions because of the students. These are new students to college and are not very comfortable with science yet. Most of them don't know how to study and they do better if there are just fact, knowledge-based questions. And, since it's an introductory course, again, I don't want them to get discouraged, so I give those questions because I don't want them to fail."*

It is evident that these instructors held core beliefs that using higher-order questions improves student learning. Their core beliefs about higher-order Bloom's questions align with literature that has shown that using exam questions that assess higher-order thinking results in students studying in ways that promote deeper, long-term, conceptual understanding (Black and Wiliam, 1998; Stanger-Hall, 2012; Jensen *et al.*, 2014). Numerous national calls to action (e.g., AAAS, 2011) suggest that instructors should craft exams that primarily assess higher levels of thinking. However, instructors' core beliefs that there are benefits to using low-level Bloom's questions on exams, conflating cognitive level with difficulty, and a fear of discouraging students by using difficult questions seemed to dissuade some instructors from testing students using mostly higher-level Bloom's questions. As a result, students who earn lower grades in a BIO101 course where they are tested with more cognitively challenging questions may have a more sophisticated understanding of biology than students who earn a higher grade in a BIO101 course with a different instructor who tests with less cognitively challenging questions.

What question format do instructors use on exams?

All BIO101 instructors that we interviewed indicated that their exams were comprised of mostly, if not exclusively, closed-ended questions (e.g., multiple choice, true/false) as opposed to open-ended questions (e.g., short answer, essay). This was one of the only commonalities in exams among all instructors of BIO101. Instructors, like Geeta, explained that this was because of a need to balance the large class sizes and limited grading support, which meant that grading open-ended questions was often not feasible.

> *Geeta: "I would say [that I use close-ended questions] mainly for convenience. These are huge classrooms. It's going to be hard for me to hand grade essays and I just didn't want to do that. [...] It's a convenience thing, yes. This way it's Scantron-based and I send it to the testing center to be scored."*

Yet, many of the instructors discussed that they would ideally administer open-ended questions and that the difference between what they wished they could do and what they actually do is because of the logistical constraint of large courses. Geeta went on to express that she would have been more likely to use open-ended questions had she been provided with adequate grading support by her institution.

> *Geeta: "I don't have the time to hand grade essays, although I would like to see some essays or some drawings, but then I'd have to hand grade everything. I don't like to give exams to undergrad teaching assistants to grade [...] I don't feel confident that they can hand grade the material. If my teaching assistant was a Ph.D. student, I would be okay with them grading it."*

This sentiment about the lack of grading support was echoed by Alex when asked why his exams were comprised entirely of closed-ended questions.

> *Alex: "The question is, who's going to grade it? [In this course], we don't have any support. [...] Open-ended questions, either for the midterm or the final – who is going to grade them? There is no support. Lack of teaching assistants."*

Additionally, some instructors highlighted that even with unlimited time, they worried about their ability to grade hundreds of essays fairly. Rater drift, or how the accuracies of one's grading can change over time, has been reported to be a problem in grading open-ended questions (Guilford, 1936; Engelhard, 1994; Longford, 1994; Clauser *et al.*, 2006; Harik *et al.*, 2009).

> *Ted: "I honestly feel that I am not capable of fairly grading more than about 30 or 40 essay questions [...] Once I get to about the 35th essay answering the same question, I don't feel like I can grade it fairly."*

Alternatively, some instructors discussed that student preference and comfort influenced why they administered closed-ended questions.

> *Hira: "Students tend to like multiple choice or matching and those kind of questions. Somehow it's not as intimidating to them."*

> *Geeta: "The students like [closed-ended questions] too. They don't like drawing and writing essays. They like multiple choice. They seem to like multiple choice so that part, I never had complaints about that in my evaluations [...] Again, I'm not sure if the students will really be interested in drawings because they've told me they're very happy with multiple-choice and matching questions. They don't want to draw. They don't want to write essays at that level at least."*

Several instructors explained that, although they felt obligated to use closed-ended questions, they worried they were compromising their ability to accurately evaluate students' conceptual understanding. Mia mentioned she used multiple-choice questions because of her large class sizes, but explicitly said she hated using multiple-choice questions.

> *Mia: "There's 450 students in the class and I don't want to grade papers. [...] It's just a numbers thing. I actually really hate multiple-choice questions because I think that students can misinterpret a question and I will never know that they're misinterpreting it. Whereas if they're doing short answers I can tell more clearly if they're misinterpreting the question. [...] But totally just a numbers thing."*

Similarly, Hira felt like closed-ended questions did not require students to think in ways that open-ended questions did, but that she could not use open-ended questions because of the challenges of grading many essays in large classrooms.

> *Hira: "I believe short answer questions actually make the students think better, more, and I can understand how. I can have a better feeling of if they understood the concept because they don't have a way to just guess or just put down a short answer, they have to really explain. I would like to include [short answer questions] on all my tests, but I could not do that because of the number of students."*

Notably, some instructors are conflating open-ended questions with higher cognitive-level questions; it is possible to write high cognitive-level closed-ended questions (Hancock, 1994). However, their decisions to strictly limit exams to primarily closed-ended questions may have consequences for student performance and student learning. Numerous studies have demonstrated that the use of closed-ended multiple-choice questions may favor the performance of certain cohorts of students (e.g., males outperform compared to females, white students outperform compared to other racial/ethnic/national identities; (Carlton and Harris, 1992; Harris and Carlton, 1993; Bastick, 2002; Lindberg *et al.*, 2010). Also, when students perceive exams will contain mostly open-ended questions, they will study in ways that focus on developing a deeper conceptual understanding of content than if they perceive exams will contain closed-ended questions (Rickards and Friedman, 1978; Thomas and Bain, 1984; Entwistle and Entwistle, 1991).

What resources do instructors use to construct exams?

While some instructors wrote some new questions for exams each time they offered BIO101, no instructor indicated they wrote an entirely new exam each time they taught the course. Instead, instructors relied heavily on questions from previously presented course material, test banks, and exam questions from previous semesters.

Several instructors indicated that when choosing exam questions, they either used exact questions or slightly changed questions that had previously been presented to students in class. Often, these questions were in the form of clicker questions during class, homework assignments, or quizzes. For example, Alex discussed that he re-used some of the questions from the course quizzes on the final exam because he had expectations that students should know those specific concepts being assessed and because the questions were readily available.

> *Alex: "There are going to be questions that are exactly the questions that I have already asked them, because it's a closed-book*

> *final. Such as the question I ask them in the quizzes. I have expectations for the students to know those."*

Time was often an important factor that instructors considered when constructing exams. For example, Hira relied on test-bank questions to save time spent writing questions. She, similar to other instructors, modified questions from test banks to make them better, but sometimes used questions that she was not completely satisfied with because she ran out of time to modify them.

> *Hira: "Some of the questions that I took from the question bank, I did not like the answers as much, but I did not have a chance to spend a lot of time to work on the questions. So that's why I don't feel 100% satisfied with my exams, but I will improve them as I teach the next time."*

Although Mia also expressed that time was an important factor in many of her decisions, she rarely used test bank questions because of her concerns that test bank questions are too simplistic.

> *Mia: "When I looked up a question in one of the test banks, the questions are too basic. I feel [students] need to be challenged a little bit with the material."*

There is a tension for instructors: they are balancing a desire for high quality questions with limited time. Mia, who had been teaching for many years, was able to avoid using test-bank questions and instead relied more on her previously used exam questions. However, Hira was a relatively new instructor of this course, so she felt she had to use test-bank questions, even though she thought they were suboptimal.

All of the BIO101 instructors indicated that they sampled from exam questions they had administered in previous semesters when constructing a new test. They usually modified the questions before using them, but sometimes they used the exact question. One of the themes that emerged as to why instructors used questions from previous exams is that they felt they had figured out the best way to write certain types of questions and did not want to try to find new ways to write those questions, as discussed by Ted.

> *Ted: "I sample from previous exams because, there's only so many ways you can write a question about a neutron. Once I've written it, I don't need to waste time trying to figure out a new way to write it. It's written. Grab it from an old exam, put it together and move on."*

Another key factor that influenced instructors' use of previous exam questions was that instructors felt that their previous exam questions were of high quality and were effective questions. For example, Mia discusses drawing on previous exam questions for this reason.

> *Mia: [I use questions from past exams] because they're good questions. Good questions make students think. I think it's [a good question] when I've seen in the past that there's a separation between how different students perform on the test. In particular, that the higher quartile performs substantially better than the lower quartile of the class."*

Although all instructors discussed drawing upon questions previously used in class, questions from prior exams, and/or questions from test banks, some instructors decided to write new exam questions. Pete writes some new questions every semester for each exam because he gives back exams to students.

> *Pete: "One big reason [that I write new questions] is because I give the tests back to students and post the keys so that they can learn from them, and I never wanted to have students that were in fraternities or sororities to have an unfair advantage because they had access to those questions."*

Pete was worried that student organizations such as fraternities and sororities have collected exams from members and created "test banks" of exams from previous years that are made available to new members of the organization. He did not want students who are involved in these organizations to have access to questions that all students in the course would not have access to, so he constructed many new questions for each exam. This is in direct contrast to Ted who uses questions from old exams and rarely writes new questions. He is aware that students may have his old exams, but is not as concerned with it anymore.

> *Ted: "As much as I've gone through these debates in my head so many times, 'Oh, my gosh, students have previous exams,' or, 'Students could have seen this question already.' Finally, I've gotten to the point where I'm like, 'You know what? I don't care. So you've seen the question already. Do you know what a neutron is? Then get the question right and let's move on.'"*

When asked why he drew on pre-existing resources, Alex brought up the challenge of the time needed to write what he perceived to be good questions with the need to balance his research and teaching commitments.

> *Alex: "It takes time to make good questions. I don't have the time to make good questions, based on how the whole system works, having to do research [and] teaching."*

The decision to write new questions or use previously used questions presents a possible tension between the integrity of an exam with the time-saving benefits of using previously administered questions that some students may have access to. This tension arose despite the evidence that indicates that students post answers to test bank questions online and share copies of previous exams (Campbell *et al.*, 2000; Shon, 2006; The Ticker, 2010). Instructors also acknowledged a tension between developing new, high-quality questions and balancing research and teaching, which may be forcing instructors to make decisions that may be suboptimal for their students. Time is the limiting factor in both sets of tensions and has been consistently described as a factor that influences college instructor decisions about teaching practices (Henderson and Dancy, 2007; Michael, 2007; Brownell and Tanner, 2012; Shortlidge *et al.*, 2016).

How do instructors make decisions about how exams are administered?

The majority of the BIO101 instructors administered exams as an in-person, paper-based exam. However, a couple of instructors administered their midterms in an online format that students could take at home. The primary reason for administering online exams was to maximize class time so that the exam could cover more material. Lawrence also administered online exams to reduce student anxiety associated with exams and potentially ask more challenging questions.

> *Lawrence: "If I [give exams] in class, then I'm restricted to just 50 minutes of the period and I cannot cover all the things I*

> *want to cover in the kind of level of questions that I want to use. [Using online exams] enables me to ask a little bit more difficult questions. I also want to erase from the question any kind of perimeters of exam anxiety."*

Alex elaborated on what he perceived to be the benefit of an "open book, open resource, online exam," indicating that such an exam is much closer to what students would be expected to do in their future careers.

> *Alex: "[An open book, open resource, online exam] is closer to the real world. When you get outside of college, the students, they have their own business, or they become scientists or physicians, they have unlimited resources. This is how they are going to operate. They're not going to operate with closed books. This is closer to what they will be doing in the future, so the exam resembles how they're going to be functioning. The score that they get is going to be more representative how successful they will be when they get outside in the real world."*

Although these exams were open-resource, students were not allowed to work with each other on the exam. Alex and Lawrence openly acknowledge that such a format is potentially conducive to cheating and took corrective actions to balance the benefits that some students received by cheating, primarily by requiring students to complete an in-person, paper-based final exam.

> *Lawrence: "I am sure that about 10% of the students are working together [on the online midterm]. I mean, it's an open book so they can surf and get the answers off the Internet if that's what they really think is going to help them, they usually are getting busted in the final. The final serves as a corrective agent for that. I think that most students are actually honest. There are definitely students who are failing and there are definitely students who are getting very low grades, so either they don't know who to copy from or more likely they're doing the exams themselves."*

> *Alex: "Still, there is cheating going around. Some students are cheating when we have it online. The final kind of counteracts that effect and kind of balances out the disadvantages of using an online exam, which has advantages but at the same time, has disadvantages, because students can work together, so it kind of counteracts the disadvantage."*

Furthermore, Lawrence took additional corrective actions by integrating open-response style questions onto his online exams to catch cheaters.

> *Lawrence: "I decided, since I caught students cheating on an open-ended homework questions, to add a few of these questions on the midterm. I was able to uncover a few cases where I was pretty sure students copied and by careful psychological type of questioning them in my office I was able to get confessions."*

Both instructors' comments highlight a key problem with implementing online exams—that students will use completely different resources depending on their social network and/or their moral integrity. Students with strong social networks would have a higher likelihood of having potential access to another student to work with on the exam. Further, a student who is in a student organization with "test files," a collection of prior exams for courses (Shon, 2006; McCabe and Bowers, 2009), would have an unfair advantage because of access to previous exams. This is a problem for students in this BIO101 course because both Alex and Lawrence pass back their exams to students in prior semesters and use past exam questions when writing new exam questions, so it is highly likely that students with access to these test files had access to very similar—if not the same—exam questions. In contrast, those students who felt it was not right to access prior exams or work with other students would have been disadvantaged by the instructors' decision to administer these types of exams.

RECOMMENDATIONS

Based on these interviews, we propose a set of recommendations for biology departments and instructors to consider when constructing exams, particularly for courses where there are multiple sections of the course that are taught by different instructors.

1. Train instructors on best practices

The instructors who were interviewed illustrated a range of familiarity with best practices for exam construction, which accounted for some of the differences in their decisions about how to construct exams. Specifically, the instructors who were associated with the discipline-based education research community held a number of beliefs that aligned with best practices, likely because they are familiar with the education literature and/or have attended training on evidence-based teaching practices (Pfund *et al.*, 2009; Yale Center for Teaching and Learning, 2018).

Importantly, there is no required pedagogical training focused on exams for college instructors at this institution. Even though peer observations are required for college instructors at some institutions, this often only consists of a classroom visit and exams are often not evaluated (Blackmore, 2005). A solution could be to familiarize instructors with best practices for writing exams that are outlined by the education literature and national recommendations. Additionally, some of the instructors in this study expressed concerns that their students may be incapable of thinking at more cognitively challenging levels, but it has been proposed that instructors should be engaging students in their own learning process in order to develop these critical thinking skills (Handelsman *et al.*, 2004).

2. Enhance exam quality and consider uniformity in exams for the same course

There is often very little oversight of exam development within departments (Laverty *et al.*, 2016). Departments can promote the creation of high quality assessments by implementing peer-review programs for exams. As part of this peer-review process, departments could work to enhance exam quality by having instructors discuss literature on best practices for designing exams. If instructors become more aware of best practices, they may begin to incorporate these practices into their exams and use these principles to help guide the peer review of other instructor's exams.

Alternatively, departments can consider creating more uniform exams for different sections of the same course by having instructors create a common exam that will be administered in all sections. Alternatively, instructors within departments can work together to create a bank of quality exam questions that they can add to and draw from throughout the semester. This way, there could be a degree of autonomy, but agreement on what constitutes an appropriate question.

3. Improve exam integrity to maximize fairness for students

While the responsibility of adhering to the values of academic integrity is often placed solely on students (Whitley and Keith-Spiege, 2012), instructors may inadvertently be promoting breaches of academic integrity by allowing exam questions to be available to some students but not others. For example, some instructors may return exams to students because research shows that students learn more when they can see their mistakes (Mason *et al.*, 2016). However, if instructors do not create exams with all new questions each semester, some students may have access to old exams while others do not, disadvantaging students who either do not have access to old exams or who have higher levels of academic integrity. Thus, if an instructor decides to give back exams, they should create new questions each semester to uphold academic integrity. Alternatively, if instructors plan to re-use exams, they may want to resort to alternative methods of letting students review exam questions (e.g., having access to exams only during office hours).

LIMITATIONS

We acknowledge that there may be sampling bias as our recruitment process relied on instructors volunteering to participate. However, the diversity of answers supports the assertion that we recruited instructors who think differently about the same phenomenon of developing exams for the same biology course. Another potential limitation is that instructor responses are self-reported and may have been influenced by social desirability bias (Zerbe and Paulhus, 1987; Grimm, 2010). Additionally, because the instructors were self-reporting on exams, there may be a disconnect between what their actual exams look like and their self-reported exams. Lastly, given that this exploratory study was limited to instructors teaching one specific course at one institution, any conclusions from this study should be interpreted as exploratory and future research should build upon this study.

ACKNOWLEDGEMENTS

We would like to thank the instructors of BIO101 for volunteering to participate in this study. Additionally, we would like to thank Dr. Sarah Dalrymple and Dr. Brian Couch for their assistance in piloting initial iterations of the interview questions.

We'd also like to thank the members of Arizona State University's (ASU) Biology Education Research Lab for their feedback and comments on the project and manuscript.

Support for this study was provided by the National Science Foundation Improving Undergraduate STEM Education program (award number 1711272), ASU's School of Life Sciences, and ASU's School of Life Sciences Undergraduate Research (SOLUR) program.

This research was conducted under the approved IRB protocol #00003837, Arizona State University.

CDW and ALH contributed equally to this work.

REFERENCES

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 Sciences Education*, *6*(2), 85–89. doi: 10.1187/cbe.07-03-0017

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

American Dental Association (ADA) (2018). Dental school admissions. Retrieved from https://www.ada.org/en/education-careers/careers-in-dentistry/be-a-dentist/applying-for-dental-school

Anderson, L., Krathwohl, D., & Bloom, B. (2001). *A Taxonomy For Learning, Teaching, And Assessing: A Revision Of Bloom's Taxonomy Of Educational Objectives*. NY: Longman.

Anderson, R. (1995). Curriculum reform: dilemmas and promise. *Phi Delta Kappan*, *77*(1), 33–36.

Association of American Medical Colleges (AAMC) (2018). Admission requirements for medical school. Retrieved from https://students-residents.aamc.org/choosing-medical-career/article/admission-requirements-medical-school/

Bastick, T. (2002). Gender differences for 6–12th grade students over Bloom's cognitive domain. Presented at the Western Psychological Association, WPA Convention, Irvine, CA.

Black, P., & Wiliam, D. (1998). Inside the black box: raising standards through classroom assessment. *Phi Delta Kappan*, *80*, 139-144. 146-148.

Blackmore, J. A. (2005). A critical evaluation of peer review via teaching observation within higher education. *International Journal of Educational Management*, *19*(3), 218–232.

Bloom, B., Englehart, M., Furst, E., Hill, W., & Krathwohl, D. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. NY: Longmans, Green.

Brownell, S. E., Hekmat-Scafe, D. S., Singla, V., Seawell, P. C., Imam, J. F., Eddy, S., … Cyert, M. S. (2015). A high-enrollment course-based undergraduate research experience improves student conceptions of scientific thinking and ability to interpret data. *CBE Life Sciences Education*, *14*(2), ar21. doi: 10.1187/cbe.14-05-0092

Brownell, S. E., Freeman, S., Wenderoth, M., & Crowe, A. (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. doi: 10.1187/cbe.13-12-0233

Brownell, S. E. & Tanner, K. (2012). Barriers to faculty pedagogical change: lack of training, time, incentives, and…tensions with professional identity? *CBE Life Sciences Education*, *11*(4), 339–346. doi: 10.1187/cbe.12-09-0163

Campbell, C., Swift, C., & Luther, D. (2000). Cheating goes hi-tech: online term paper mills. *Journal of Management Education*, *24*(6), 726–740.

Carlton, S. T., & Harris, A. M. (1992). Characteristics associated with differential item functioning on the scholastic aptitude test: gender and majority/minority group comparisons. *ETS Research Report Series, 1992*(2), i-143. doi: 10.1002/j.2333-8504.1992.tb01495.x

Clauser, B. E., Harik, P., & Margolis, M. J. (2006). A multivariate generalizability analysis of data from a performance assessment of physicians' clinical skills. *Journal of Educational Measurement, 43(3)*, 173–191. doi: 10.1111/j.1745-3984.2006.00012.x

Cooper, K. M., Soneral, P. A. G., & Brownell, S. E. (2017). Define your goals before you design a cure: a call to use backward design in planning course-based undergraduate research experiences. *Journal of Microbiology & Biology Education*, *18*(2). doi: 10.1128/jmbe.v18i2.1287

Creech, L. R., & Sweeder, R. D. (2012). Analysis of student performance in large-enrollment life science courses. *CBE Life Sciences Education*, *11*(4), 386–391. doi: 10.1187/cbe.12-02-0019

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. doi: 10.1187/cbe.08-05-0024

Eddy, S. L., Brownell, S. E., & Wenderoth, M. P. (2014). Gender gaps in achievement and participation in multiple introductory biology classrooms. *CBE Life Sciences Education*, *13*(3), 478–492. doi: 10.1187/cbe.13-10-0204

Eddy, S. L. & Brownell, S. E. (2016). Beneath the numbers: a review of gender disparities in undergraduate education across science, technology, engineering, and math disciplines. *Physical Review Special Topics - Physics Education Research*, *12(2) [020106]*. doi: 10.1103/PhysRevPhysEducRes.12.020106

Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93–112. doi: 10.1111/j.1745-3984.1994.tb00436.x

Entwistle, N. J., & Entwistle, A. (1991). Contrasting forms of understanding for degree examinations: the student experience and its implications. *Higher Education*, *22*(3), 205–227. doi: 10.1007/BF00132288

Freeman, S., O'Connor, E., Parks, J. W., Cunningham, M., Hurley, D., Haak, D., … & Wenderoth, M. P. (2007). Prescribed active learning increases performancec in introductory biology. *CBE Life Sciences Education*, *6(2)*, 132–139. doi: 10.1187/cbe.06-09-0194

Glaser, B., & Strauss, A. (2009). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. New Brunswick, NJ: Transaction.

Grimm, P. (2010). Social desirability bias. *Wiley International Encyclopedia of Marketing*, *2*. doi: 10.1002/9781444316568.wiem02057

Guilford, J. (1936). *Psychometric Methods*. NY: McGraw-Hill.

Hancock, G. R. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *Journal of Experimental Education*, *62*(2), 143–157.

Handelsman, J., Ebert-May, D., Beichner, R., Bruns, P., Chang, A., DeHaan, R., … & Wood, W. B. (2004). Scientific teaching, *304*(5670), 521+.

Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., Swanson, D., & Nandakumar, R. (2009). An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement*, *46*(1), 43–58. doi: 10.1111/j.1745-3984.2009.01068.x

Harris, A. M., & Carlton, S. T. (1993). Patterns of gender differences on mathematics items on the Scholastic Aptitude Test. *Applied Measurement in Education*, *6*(2), 137–151. doi: 10.1207/s15324818ame0602_3

Henderson, C., & Dancy, M. H. (2007). Barriers to the use of research-based instructional strategies: the influence of both individual and situational characteristics. *Physical Review Special Topics - Physics Education Research*, *3(2) [020102]*. doi: 10.1103/PhysRevSTPER.3.020102

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. doi: 10.1007/s10648-013-9248-9

Katz, L. G., & Rath, J. (1992). Six dilemmas in teacher education. *Journal of Teacher Education*, *43*(5), 376–385. doi: 10.1177/0022487192043005007

Kearney, M. H. (2001). New directions in grounded formal theory. *Using Grounded Theory in Nursing*. Schreiber, R.S., Stern, P.N. (Eds.). NY: Springer.

Kendig, K. (2013). How college grading is different from high school. *USA Today*. Retrieved from http://college.usatoday.com/2013/05/31/how-college-grading-is-different-from-high-school/

Kost, L. E., Pollock, S. J., & Finkelstein, N. D. (2009). Characterizing the gender gap in introductory physics. *Phys. Rev. ST Phys. Educ. Res.*, *5*(1). doi: 10.1103/PhysRevSTPER.5.010101

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. doi: 10.1371/journal.pone.0162333

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: a meta-analysis. *Psychological Bulletin*, *136*(6), 1123–1135. doi: 10.1037/a0021276

Longford. (1994). Reliability of essay rating and score adjustment. *Journal of Educational and Behavioral Statistics*, *19(3)*, 171–200. doi: 10.1002/j.2333-8504.1993.tb01563.x

Mason, A., Yerushalmi, E., Cohen, E., & Singh, C. (2016). Learning from mistakes: the effect of students' written self-diagnoses on subsequent problem solving. *The Physics Teacher*, *54*(87), 87. doi: 10.1119/1.4940171

McCabe, D. L., & Bowers, W. J. (2009). The relationship between student cheating and college fraternity or sorority membership. *NASPA Journal*, *46*(4), 573–586.

Michael, J. (2007). Faculty Perceptions about barriers to active learning. *College Teaching*, *55*(2), 42–47.

Miyake, A., Kost-Smith, L. E., Finkelstein, N. D., Pollock, S. J., Cohen, G. L., & Ito, T. A. (2010). Reducing the gender achievement gap in college science: a classroom study of values affirmation. *Science*, *330*(6008), 1234–1237. doi: 10.1126/science.1195996

Nespor, J. (1990). Grades and knowledge in undergraduate education. *Journal Of Curriculum Studies*, 22(6), 545-556. doi: 10.1080/0022027900220603

Pfund, C., Miller, S., Brenner, K., Bruns, P., Chang, A., Ebert-May D., … Handelsman, J. (2009). Summer institute to improve university science teaching. *Science*, *324*(5926), 470-471. doi: 10.1126/science.1170015

Rauschenberger, M. M., & Sweeder, R. D. (2010). Gender performance differences in biochemistry. *Biochemistry and Molecular Biology Education*, *38(6)*, 380–384. doi: 10.1002/bmb.20448

Rickards, J. P., & Friedman, F. (1978). The encoding versus the external storage hypothesis in note taking. *Contemporary Educational Psychology*, *3*(2), 136–143. doi: 10.1016/0361-476X(78)90020-6

Roediger, H., Agarwal, P., Kang, S., & Marsh, E. (2009). Benefits of testing memory. In Davies, G. M., and Wright, D. B. (Eds.), *Current Issues in Applied Memory Research*. NY: Psychology Press.

Roediger, H., Putnam, A., & Smith, M. (2011). Ten benefits of testing and their applications to educational practice. In Mestre, J.P., and Ross, B. H. (Eds.), *Cognition in Education*. San Diego, CA: Elsevier/Academic Press

Shon, P. (2006). *How College Students Cheat on In-class Examinations: Creativity, Strain, and Techniques of Innovation*. Ann Arbor, MI: MPublishing, University of Michigan Library, 1.

Shortlidge, E. E., Bangera, G., & Brownell, S. E. (2016). Faculty perspectives on developing and teaching course-based undergraduate research experiences. *BioScience*, *66*(1), 54–62. doi: 10.1093/biosci/biv167

Smith, G. (2007). How does student performance on formative assessments relate to learning assessed by exams. *Journal of College Science Teaching*, *36(7)*, 28–34.

Smith, J. I., Combs, E. D., Nagami, P. H., Alto, V. M., Goh, H. G., Gourdet, M. A. A., … Tanner, K. D. (2013). Development of the biology card sorting task to measure conceptual expertise in biology. *CBE Life Sciences Education*, *12*(4), 628–644. doi: 10.1187/cbe.13-05-0096

Stanger-Hall, K. F. (2012). Multiple-choice exams: an obstacle for higher-level thinking in introductory science classes. *CBE Life Sciences Education*, *11*(3), 294–306. doi: 10.1187/cbe.11-11-0100

Stemler, S. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9*(4), 1-11.

Strauss, A., & Corbin, J. (1990). *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. London: Sage..

Tai, R., & Sadler, P. (2001). Gender differences in introductory undergraduate physics performance: university physics versus college physics in the USA. *International Journal of Science Education*, *23(10)*, 1017–1037.

The Ticker. (2010). Cheating on University of Central Florida test was aided by use of textbook questions. *The Chronicle for Higher Education news blog*. Retrieved from

Thomas, P., & Bain, J. (1984). Contextual dependence of learning approaches: the effects of assessments. *Human Learning*, *3(4)*, 227–240.

University of Arizona (2016). Introductory biology home. Retrieved from https://blc.arizona.edu/introbio/

University of California, Los Angeles (2017). UCLA Registrar's office, life science courses. Retrieved from http://catalog.registrar.ucla.edu/ucla-catalog2017-585.html

Walvoord, B.E. (2010). *Assessment Clear and Simple: A Practical Guide for Institutions, Departments, and General Education*. San Francisco, CA: Jossey-Bass.

Whitley, B. E., Jr, & Keith-Spiege, P. (2012). *Academic Dishonesty: An Educator's Guide*. NY: Lawrence Erlbaum Associates.

Wiggins, G., & McTighe, J. (1998). *Understanding by Design*. Alexandria, VA: Association for Supervision and Curriculum Development.

Wright, C. W., Eddy, S. L., Wenderoth, M. P., Abshire, E., Blankenbiller, M., & Brownell, S. E. (2016). Cognitive difficulty and format of exams predicts gender and socioeconomic gaps in exam performance of students in introductory biology courses. *CBE Life Sciences Education*, *15*(2), ar23. doi: 10.1187/cbe.15-12-0246

Yale Center for Teaching and Learning (2018). Summer institutes on scientific teaching. Retrieved from https://www.summerinstitutes.org/

Zerbe, W. J., & Paulhus, D. L. (1987). Socially desirable responding in organizational behavior: a reconception. *Academy of Management Review*, *12*(2), 250–264. doi: 10.5465/AMR.1987.4307820

Supplemental Materials

Table S1. Questions asked to participants during semi-structured interview, presented in the order in which they were asked.

Describe the entire process by which you go through when constructing a typical exam for the course you indicated on the survey you submitted.
You characterized your typical exams as containing ___% close-ended questions and ___% of questions that were open-ended questions. Why did you design your typical exams in that way?
You characterized your typical exam as containing ___% questions that test knowledge of definitions, memorization or facts, and/or descriptions of processes while the remaining ___% tests the ability to synthesize, analyze, evaluate, and/or apply knowledge. Why did you design your typical exams in this way?
Consider a typical exam, what are the benefits and costs to writing and implementing a typical exam and why do you consider these to be benefits and/or costs?
What are your goal(s) when writing and implementing a typical exam and why are these your goal or goals?
Did the goal(s) you have when writing and implementing your typical exams align with your long-term career teaching or personal goal(s)? If so, how and why?
After you administer your exam, how do you use that exam and why would you use that exam in that way?
Does the way you use exams align with your long-term career in teaching or personal goal(s)? How and why?
What are the barriers you experienced when writing and implementing a typical exam and why do you consider these to be barriers?
Is it difficult for you to write your exams? If so, why or why not?
How much effort do you put into constructing a typical exam and why that level of effort?

Table S2. Questions asked to participants during the online survey administered prior to the interview, presented in the order in which they were asked. Unless otherwise indicated, responses were open-ended.

What is your percentage breakdown for research, teaching, service, and/or other (e.g., administrative), as assigned by your employer?
How many semesters have you taught BIO101?
Approximately how many students were in your most recent iteration of BIO101?
In your most recent iteration of BIO101, did you teach this course with another instructor that was not a TA? Yes or No.
In your most recent iteration of BIO101, did you and your co-instructor collaborate when writing exams? Yes or No.
In your most recent iteration of BIO101, did you have teaching assistant(s) (TA's) that assisted in writing exam questions? Yes or No.
In your most recent iteration of BIO101, did you have teaching assistant(s) (TA's) that assisted in grading exams? Yes or No.
In your most recent iteration of BIO101, what percentage of the overall course grade did all of the exams account for?
In your most recent iteration of BIO101, approximately how many questions were on a typical exam you wrote?
In your most recent iteration of BIO101, approximately what percentage of the questions on a typical exam you wrote were closed-ended questions (e.g., multiple choice, true false, matching.) versus open-ended questions (e.g., short answer, essay, fill in the blank, graphing)?
In your most recent iteration of BIO101, approximately what percentage of the questions on a typical exam you wrote assessed students' knowledge of definitions, facts, and/or descriptions of processes versus students' ability to synthesize information, analyze information/data, evaluate information/data, and/or apply their knowledge to new situations?
In your most recent iteration of BIO101, approximately what percentage of the questions on a typical exam you wrote contained material that had been assessed on a previous exam(s) versus material that had not been previously assessed on an exam(s)?
In your most recent iteration of BIO101, when constructing your exams did you use questions that were taken from external resources (e.g., textbook question banks)? Yes or No. If so, please describe what resources you have used.
In your most recent iteration of BIO101, did you reuse any old exam questions? Yes or No.
In your most recent iteration of BIO101, did anyone else proof read the exams you wrote prior to administering the exams to students? Yes or No.
In your most recent iteration of BIO101, did you post exam keys and/or allow students to keep a copy of the exams? Yes or No.

Rubric S3. Rubric used to code decisions instructors indicated they made during the semi-structured interview.

Content to test

1. Learning outcomes: *When an instructor discusses using learning objectives/outcomes and/or goal when constructing exams.*

2. Lectures: *When an instructor discusses using PowerPoints, lectures, and/or presentations when constructing exams.*

3. Coverage of content: *When an instructor discusses that they try to ensure even coverage (or uneven coverage) of content from a lecture or a chapter on exam.*

4. Number of concepts being tested: *When an instructor discusses whether they have a large amount of different concepts being covered or fewer concepts on an exam.*

5. Depth of content: *When an instructor discusses the depth of content they put on an exam, including which concepts are more important than others to assess. This includes when they decide to include fundamental concepts in the course for students to learn or focus on specific definitions.*

Question characteristics

1. Open-ended questions vs. close-ended questions

   Open-ended questions: *When students are forced to come up with their own answer (i.e. short answer, essay, fill in the blank).*

   Closed-ended questions: *When students are drawing on predetermined answer choices (i.e. true/false, multiple choice, matching).*

2. Low-level Bloom's questions vs. high-level Bloom's questions

   Lower-level questions: *Definitions, fact based, memorization, assessing specific details, things (e.g. questions) that they have seen before, and/or descriptions of processes.*

   Higher-level questions: *Synthesis, analysis, evaluation, critical thinking, assessing big ideas, and/or application (including novel questions/scenarios).*

Materials used to construct the exam

1. Past exams questions: *When an instructor discusses any exam question that was pulled or drawn on from a previous exam within the same semester or in past semesters or years within the courses that this instructor has taught.*

2. Textbook test bank questions: *When an instructor discusses using test bank questions or textbook test bank questions*

3. Questions previously presented to students: *When an instructor discusses using questions previously presented to students in class, such as clicker questions, quizzes, homework, etc.*

4. Writing brand new questions: *When an instructor discusses the decision to write entirely new questions from scratch.*

Assessment format and delivery

1. Paper exams – *When an instructor discusses using a paper-based in-person exam (e.g. traditional exams, Scantron).*

2. Online exams – *When an instructor discusses using an online out-of-class exam.*

Accessing Materials

No additional materials available online.