Teacher Knowledge for Active-Learning Instruction: Expert–Novice Comparison Reveals Differences

A. J. Auerbach,[†] M. Higgins,[†] P. Brickman,[‡] and T. C. Andrews[†]*
[†]Department of Genetics and [‡]Department of Plant Biology, University of Georgia, Athens, GA 30602

ABSTRACT

Active-learning strategies *can* improve science, technology, engineering, and mathematics (STEM) undergraduates' abilities to learn fundamental concepts and skills. However, the results instructors achieve vary substantially. One explanation for this is that instructors commonly implement active learning differently than intended. An important factor affecting how instructors implement active learning is knowledge of teaching and learning. We aimed to discover knowledge that is important to effective active learning in large undergraduate courses. We developed a lesson-analysis instrument to elicit teacher knowledge, drawing on the theoretical construct of teacher noticing. We compared the knowledge used by expert ($n = 14$) and novice ($n = 29$) active-learning instructors as they analyzed lessons. Experts and novices differed in what they noticed, with experts more commonly considering how instructors hold students accountable, topic-specific student difficulties, whether the instructor elicited and responded to student thinking, and opportunities students had to generate their own ideas and work. Experts were also better able to support their lesson analyses with reasoning. This work provides foundational knowledge for the future design of preparation and support for instructors adopting active learning. Improving teacher knowledge will improve the implementation of active learning, which will be necessary to widely realize the potential benefits of active learning in undergraduate STEM.

INTRODUCTION

Calls for incorporating active-learning instruction in undergraduate science, technology, engineering, and mathematics (STEM) courses abound and have become increasingly strident as evidence continues to accumulate that active learning can be much more effective than traditional lecture. Some have even argued that it is "unethical" to teach any other way (Waldrop, 2015). However, the conversation often does not proceed to address how we will prepare and support instructors to develop knowledge and skills necessary to *effectively* implement these strategies. Yet the results instructors achieve implementing active-learning instruction vary substantially (e.g., Pollock and Finkelstein, 2008; Andrews *et al.*, 2011). Instructors commonly implement active-learning strategies differently than intended, making decisions to omit parts that are crucial to student learning (Henderson and Dancy, 2009; Dancy *et al.*, 2016; Turpen and Finkelstein, 2009). Therefore, it may be unreasonable to expect that instructors will immediately achieve impressive learning gains using active learning. The danger of promising—implicitly or explicitly—that active learning is much more effective than traditional lecture (without caveat) is that instructors will try it, achieve mediocre results, and quit—perhaps with diminished faith in the value of education research.

We will likely be more successful in engendering widespread instructional change in undergraduate STEM if we recognize that active-learning instruction differs in important ways from traditional approaches and that the knowledge a teacher uses influences how he or she plans and implements active learning (e.g., Park *et al.*, 2011;

Jay Labov, *Monitoring Editor*
Submitted August 11, 2017; Revised November 13, 2017; Accepted December 8, 2017
CBE Life Sci Educ March 1, 2018 17:ar12
DOI:10.1187/cbe.17-07-0149
*Address correspondence to: T. C. Andrews (tandrews@uga.edu).

Santagata and Yeh, 2014; Stains and Vickrey, 2017). Importantly, an instructor's knowledge of teaching and learning—not just content knowledge—affects student outcomes (e.g., Hill *et al*., 2005; Sadler *et al*., 2013; Blömeke *et al*., 2015). Therefore, there is a critical need to determine what teacher knowledge is important for effective implementation of active-learning instruction in undergraduate STEM. Not meeting this need limits our ability to create evidence-based preparation and support for instructors to achieve the benefits of active learning for their students.

Teacher noticing is a theoretical construct of teacher expertise that has been productive for investigating knowledge among K–12 instructors (Kersting, 2008; Talanquer *et al.*, 2012). Teacher noticing recognizes that classrooms are complex social environments, and that teachers cannot attend to everything occurring in a classroom (Sherin and van Es, 2008). "Noticing" is knowledge in action that teachers use to reason about what they see and make instructional decisions in real time (Sherin *et al*., 2011). Doing this efficiently requires rich knowledge built through experience (van Es, 2011). The ability of teachers to notice and reason about events in the classroom predicts instructional quality and student learning (van Es and Sherin, 2008; Kersting *et al*., 2012; Santagata and Yeh, 2014).

More experienced K–12 teachers notice differently than novice teachers, drawing on knowledge that newer teachers lack to critically analyze teaching and learning in a classroom. More experienced teachers pay more attention to student thinking and to the relationship between teaching strategies and student thinking (van Es and Sherin, 2008; van Es, 2011; Kisa and Stein, 2015). They are better able to make inferences about what is happening that go beyond what is observable (Santagata and Angelici, 2010; Kisa and Stein, 2015) and can support these inferences with reasoning, making connections between events they observe and principles of teaching and learning (van Es, 2011). In contrast, novice teachers primarily describe what they notice without reasoning (van Es, 2011). Experienced teachers are also able to make a greater number of suggestions and more detailed suggestions for improving instruction (Santagata and Angelici, 2010; Kersting *et al*., 2012). Making suggestions requires recognizing what is missing and proposing solutions.

Teacher noticing has not yet been used as a theoretical construct for active-learning instruction in undergraduate STEM, but the construct is well suited to this context. Teacher noticing emphasizes the importance of attending to student thinking and focuses on real-time reasoning and responding in the classroom. While a lecturer can prepare an entire lesson before a class meeting and enact it without much deviation, an active-learning instructor engages students in examining and articulating their own thinking during a lesson and must be able to recognize, make sense of, and respond to student thinking in real time (Wagner *et al*., 2007).

The objective of this study was to examine teacher noticing by expert active-learning instructors and to contrast this with noticing by novice active-learning instructors. What instructors notice reveals teacher knowledge upon which they are drawing. Their reasoning about what they notice provides insight about the depth of that knowledge. We were especially interested in knowledge important for active learning in large courses, which we define as 50 or more students, because large courses present unique challenges and are common in undergraduate STEM education (National Research Council, 2012). An expert–novice approach can elucidate 1) knowledge used by experts, which can then serve as learning objectives for teaching professional development for other instructors; and 2) knowledge used by novices that can be productively built upon in teaching professional development. We investigated teacher knowledge using a lesson-analysis survey in which participants watch videos of authentic lessons and analyze the lessons. Lesson analysis is an accepted method for studying teacher noticing (e.g., Kersting, 2008; Santagata and Yeh, 2014).

We also grounded this study in a theoretical framework for distinguishing among different types of active learning, the ICAP framework. ICAP deconstructs active learning by defining four modes of cognitive engagement based on the behaviors of students as they interact with instructional materials (Chi and Wylie, 2014):

- **I**nteractive: learners collaborate with peers to generate work that goes beyond what was explicitly presented in instructional materials (e.g., building consensus).
- **C**onstructive: learners generate work that goes beyond what has been presented in instructional materials (e.g., providing an explanation, solving a problem, self-explaining).
- **A**ctive: learners make physical manipulations without adding new knowledge (e.g., rolling a die).
- **P**assive: learners receive information (e.g., listening).
- Interactive and constructive modes are "generative," because students generate their own ideas and work, and extensive empirical studies indicate that constructive and interactive active learning lead to greater learning gains than being active or passive: I > C >> A > P (Chi, 2009; Chi *et al*., 2017; Menekse *et al*., 2013).

This study addressed two research questions. Addressing the first question was important, because it provided insight into what components of instruction are more commonly noticed by experts and therefore may be especially critical for effective active-learning instruction. Answering the second question was important, because individuals with a better understanding of *why* a component of instruction is important are more likely to be able to plan and implement effective instruction themselves (e.g., Rogers, 2003; Kersting *et al*., 2012).

> Research question 1: How do experts and novices differ in *what* they notice as they analyze active-learning lessons?
> Research question 2: How do expert and novice active-learning instructors differ in their ability to *use reasoning* to support their evaluations of and suggestions for improving active-learning lessons?

METHODS

Participants

Participants included undergraduate biology instructors who met inclusion criteria of either an "expert" or "novice" active-learning instructor and taught large (50+ students) biology courses using active learning. All participants had earned a PhD in a life sciences discipline and had previously taught large college biology courses. Based on an extensive review of studies of expert and novice teachers, Palmer *et al*. (2005) recommend that researchers apply consistent criteria to distinguish experts

from novices. We considered three criteria when identifying experts and novices:

A. Years of experience using active-learning strategies in large college biology courses;
B. Evidence of effectiveness at facilitating student learning using active learning in a large biology course; and
C. Evidence of purposeful reflective practice to improve student learning in a large biology course.

Experts had four or more years of experience using active learning in large courses and met criteria B, C, or both. Novices had four or fewer years of experience and met neither criteria B nor C. Additionally, no novices had published discipline-based education research (DBER). Novices were not excluded for presenting about teaching at a professional conference or publishing lessons, but most had not done so. Each criterion is described in more detail later in this section.

We triangulated multiple data sources to confirm expert and novice criteria were met. We conducted screening interviews of all potential participants to learn about their use of active-learning strategies, how long they have been using these strategies, their typical active-learning course size, and how they approached reflection and improvement in their teaching (interview protocol in Appendix A in the Supplemental Material). We recorded, transcribed, and systematically analyzed interviews for evidence of purposeful, reflective teaching practice. We also collected CVs and used public websites to determine whether potential novices had published DBER.

Years of Experience (Criterion A). We considered years of experience important, because teacher knowledge is primarily built through iterative experience and reflection (McAlpine and Weston, 2000; McAlpine *et al*., 2006).

Evidence of Effectiveness (Criterion B). We operationalized effectiveness as student learning gains. We considered evidence of student learning gains important, because improving student outcomes is the ultimate goal of active learning. Additionally, there is wide variability in the learning gains instructors are able to achieve using active learning (e.g., Andrews *et al*., 2011). We sought instructors who had measured learning using a pre/posttest design. Experts had assessed student learning using instruments such as the Conceptual Inventory of Natural Selection (CINS; Anderson *et al*., 2002), the Introductory Molecular and Cell Biology Assessment (Shi *et al.*, 2010), and the Genetics Concept Assessment (Smith *et al*., 2008). Some instructors had used selected questions from these instruments that were aligned with their learning objectives, along with items they had created and refined over multiple semesters. Instructors had collected these data for their own purposes. We asked potential participants who had collected this type of data whether they would be willing to share their data with us, following de-identification.

We calculated learning gains as effect size. Effect size quantifies the magnitude of change in student knowledge from the time of the pretest to the time of the posttest (Middlemis *et al*., 2013). We calculated effect size using a variant of Cohen's *d* that accounts for the fact that correlations between students' pre- and posttest scores can lead to substantial overestimation of effect size (Dunlap *et al*., 1996). We considered assessment data to be sufficient evidence of effectiveness if effect size (*d*) was greater than 0.80. This cutoff is justified for several reasons. First, 0.80 is widely considered to be a large effect size (e.g., Middlemis *et al*., 2013). Second, 0.80 is large when compared with effect sizes commonly found in studies comparing pre- and posttest scores in college biology courses (e.g., Andrews *et al*., 2011). For example, only five courses (15%) achieved effect sizes greater than 0.80 on the CINS in a national study of student learning in college biology courses (Andrews *et al*., 2011).

Our standard for demonstrating effectiveness was rigorous, but it also has several limitations. First, instructors were almost always pre/posttesting for a single topic. They may be highly effective at teaching that topic and much less effective at teaching other topics. Therefore, systematically assessing student learning or student thinking across a course, rather than just for a single topic, is also important. Second, some instructors do not teach courses that include topics covered in research-based instruments, potentially excluding highly effective active-learning instructors. Third, many important learning objectives are not easily assessed with instruments that instructors can conveniently administer and score, including many core competencies outlined in *Vision and Change* (American Association for the Advancement of Science [AAAS], 2011). Fourth, pre/posttesting is not a standard practice in undergraduate biology education. Criterion C was designed to address some of these limitations.

Purposeful Reflective Practice (Criterion C). Effective teachers engage in regular, purposeful reflection on their teaching (Schön, 1987; Kane *et al*., 2004; Chan and Yung, 2015). Reflection involves monitoring student thinking or student learning; comparing these data with intended learning objectives; and making decisions to maintain, initiate, adjust, or terminate a teaching approach (McAlpine and Weston, 2000). Critical reflection on teaching experiences is paramount to developing teaching expertise and improving effectiveness (McAlpine and Weston, 2000; Kane *et al*., 2004). This criterion was met if a potential participant described, in detail, a systematic approach he or she used to learn about student thinking on a regular basis, as well as the changes he or she had made to the active-learning instruction based on what was learned through this systematic approach. These were not onetime measurements instructors made, but rather a consistent approach taken in monitoring their own effectiveness and making changes based on what they learned. We gathered evidence to determine whether an instructor engaged in purposeful reflective practice using the screening interview (Appendix A in the Supplemental Material).

Participant Identification and Recruitment

We used three approaches to identify potential participants. One way to find people using active learning in large courses, and especially people who rigorously assess their own effectiveness, is to contact people interested in biology education research. We sent a query to the Society for the Advancement of Biology Education Research (SABER) listserv asking them to help us identify experienced and inexperienced active-learning instructors. We contacted SABER members, because we expected this group to include individuals who conduct education research in their own and others' classrooms and individuals who have led teaching professional development. These

experiences provide opportunities to meet active-learning instructors with varying levels of experience. Additionally, most members of SABER have positions in life science departments and may regularly interact with colleagues about teaching (e.g., Andrews *et al*., 2016), providing additional opportunities to become aware of who is using active learning. In our email to the listserv, we asked for help identifying 1) instructors who were relatively new to using active-learning in large college biology courses and 2) instructors who had been using active-learning in large college biology courses AND have collected data to assess student learning during that time. We included a link to an online survey where listserv members could easily share this information. We identified and contacted 141 potential participants using this approach.

Another approach we used to identify experts and novices was to find initiatives across the country that aimed to improve undergraduate biology education. The National Science Foundation (NSF) has funded such projects through current and former programs. All funded projects are publicly available in a searchable list on the NSF website. We identified principal investigators (PIs) for projects funded through Widening Implementation and Demonstration of Evidence-Based Reforms and Improving Undergraduate STEM Education (IUSE) grants. We also identified PIs funded through older programs who were still active, as evidenced by their attendance at the 2016 Envisioning the Future of STEM Education conference hosted by NSF and the AAAS. These PIs could have been funded through prior programs such as Transforming Undergraduate Education in STEM or Course, Curriculum, and Laboratory Improvement. We used publicly available project descriptions to find projects related to undergraduate biology and active learning. We asked individual PIs about their own experiences using active learning and about instructors they had worked with as part of their projects. We identified and contacted 145 PIs using this approach.

Our last approach was to contact relevant organizations that offer professional development, resources, and support to college biology instructors who use or are interested in using active learning. In some cases, organizers were able to provide contact information for former participants of teaching professional development. In other cases, we used publicly available lists of participants or members to collect contact information for potential participants. We identified and contacted 55 potential participants using this approach.

We contacted each potential participant by email, briefly explained the overall purpose of the study, and asked whether he or she was available for a short (<10 minute) phone call to conduct our screening interview. We sent up to four follow-up emails to schedule this interview. We scheduled phone calls or virtual meetings to conduct screening interviews with all potential participants who responded to our emails. We asked potential participants who seemed likely to meet our expert or notice criteria to participate in the study via an online survey. Later, we systematically analyzed screening interviews for evidence of reflective practice and teaching experience and examined CVs. This further reduced the pool of participants who met our final expert and novice criteria. The online survey included lesson-analysis questions and questions about teaching practices and relevant professional experiences.

Designing and Refining a Lesson-Analysis Survey to Elicit Teacher Noticing

We iteratively developed and refined a lesson-analysis survey to elicit teacher knowledge. We modeled this instrument after prior instruments used with K–12 teachers (e.g., Kersting, 2008; Kersting *et al.*, 2012). The survey used videos of authentic active-learning lessons in large (50+ students) undergraduate biology courses as stimuli, followed by writing prompts that asked participants to evaluate the lesson and make suggestions for improvement. We refined initial versions of the survey by collecting and analyzing responses from instructors with varying levels of teaching experience and expertise and gathering expert feedback. Step-by-step details of this process are in Appendix B in the Supplemental Material.

The final version of the survey included three videos that were no more than 5 minutes long. After the first two videos, instructors responded to two written prompts:

1. What was effective and why did you think it was effective? Please use complete sentences.
2. What needs to be improved and why? How would you do it differently? Please use complete sentences.

After the third video, we asked instructors to respond to question 1. See the full text of the survey in Appendix C in the Supplemental Material.

Other Survey Questions

The online survey also asked about teaching practices and other relevant professional experiences. We measured self-reported teaching practices to test for differences in use of evidence-based teaching practices between experts and novices. We used a section of the Teaching Practices Inventory (TPI, section 3; Wieman and Gilbert, 2014). The TPI measures the extent to which college STEM instructors use research-based teaching practices. Section 3 focuses on in-class features and activities and includes questions about what the instructor and students do in class (see all questions in Appendix C in the Supplemental Material).

One concern with this approach is that it relies on instructors to accurately represent (i.e., self-report) their teaching. Smith and colleagues (2014) compared instructor's reported teaching practices on the TPI to practices seen by observers in the classroom. They found a significant negative correlation ($r = -0.509$, $p < 0.05$) between scores on Section 3 of the TPI and the percent of time that instructors were observed to be presenting information to students (e.g., lecturing, real-time writing, showing demonstrations or videos), indicating that instructors were aware of their practices (Smith *et al.*, 2014).

We also surveyed participants about relevant prior professional experiences. We asked whether participants had participated in 40 or more hours of teaching professional development and whether they had led teaching professional development. All participants were offered a $25 gift card as an incentive for survey completion. Participants who provided de-identified data demonstrating student learning gains were offered an additional $50 incentive. This study received Institutional Review Board approval at the University of Georgia before data collection, under protocol #00002116.

Data Analysis

We conducted qualitative content analysis of participants' written responses to the lesson-analysis survey to determine what they noticed and whether they reasoned about what they noticed. Qualitative analyses produced the raw data with which we made quantitative comparisons between expert and novice active-learning instructors. We used Atlas.ti to organize qualitative analyses and R for all quantitative comparisons.

Qualitative Analysis. We started by identifying what participants noticed when they analyzed active-learning lessons (research question 1). This was done in two phases. The goal of the first phase was to catalogue everything participants noticed. We generated an a priori list of codes when analyzing pilot data collected as we refined the lesson-analysis survey (Appendix B in the Supplemental Material). Each code was a short phrase we used to describe the essence of an idea. Coding began by reading participants' responses, identifying sections of text that communicated an idea, and assigning code(s) that captured the idea. When a response contained an idea we had not yet encountered in our data, we created a new code. Creating a code involved naming and defining the code. The definitions of codes evolved over time as examples accumulated. Definitions included the boundaries of the code and the breadth of ideas within the code. We continued to read participants' responses and refined our codes as appropriate.

We regularly read all the quotes within a code and within closely related codes to determine the boundaries between codes. For example, when it became clear that there were distinct categories within a code, we would split the code into two or more codes and describe their differences in detail. Alternatively, when it became clear that we were struggling to distinguish between two codes, they were combined into a single code. The complete list of codes and their definitions is called a codebook. As decisions about adding, splitting, or dividing codes were made, we revisited all previously coded responses to apply the refined codebook. Our qualitative analyses were highly iterative and collaborative. We always coded in teams of two to four researchers and discussed until we reached consensus.

At the end of the first phase of qualitative analysis, we had documented every idea that came up when participants analyzed lessons, but not all codes provided meaningful insight into teacher knowledge for active learning. Some codes had been used for quotes that seemed to be about a specific idea but were too vague for us to be sure. We did not analyze these further, because we could not be confident about the thinking underlying the statements. We also set aside a code for quotes about the instructor explaining or lecturing, as these do not constitute active learning. All other codes continued on to our next phase of qualitative analysis.

The second phase of qualitative analysis for research question 1 aimed to group the remaining codes into themes. The number of codes created to fully characterize the data was too large to be meaningfully interpreted. Grouping codes conceptually into themes made them interpretable and allowed for quantitative comparisons. We used several approaches to inform how we grouped codes into themes. First, we turned to relevant literature to inform our thinking about conceptual relationships among codes, including literature on active-learning instruction (e.g., Eddy *et al.*, 2015), student motivation (e.g., Glynn *et al.*, 2009), cognitive science (e.g., Vermunt, 1996), and course climate (e.g., Ambrose *et al.*, 2010). Second, we examined co-occurrences of codes within the same body of text to identify codes that almost always occurred together. Third, on multiple occasions three researchers involved throughout the coding process (A.J.A., M.H., and T.C.A.) independently grouped codes into themes, presented the themes to one another, and discussed their thinking to reach consensus. Fourth, we presented codes and tentative themes to other discipline-based education researchers and sought feedback. We used the insight gleaned from these approaches and repeatedly reread all quotes from all codes within a theme to make final decisions. Our final organization of codes includes 13 distinct themes that represented 30 codes. Hereafter, we refer to these themes as "components of instruction" to which participants attended in their lesson analyses. We named these components according to the goals instructors are trying to achieve.

We also used qualitative analysis to address research question 2. We determined which statements were evaluations and which were suggestions and determined when participants provided reasoning and when they did not. Our qualitative analysis for research question 1 had divided the data for each participant into "idea units," which are sections of text that address a distinct idea (Kisa and Stein 2015). We considered each idea unit to determine whether it was a suggestion or not. We considered any statement that was not a suggestion to be an evaluation, because the writing prompts to which participants responded were evaluative in nature. We next determined whether each evaluation statement and suggestion provided reasoning. Table 1 gives examples of evaluations and suggestions with and without reasoning.

Quantitative Analysis. The goal of our quantitative analysis was to compare the noticing employed by experts and novices in analyzing active-learning lessons. We analyzed all quantitative data as counts. Specifically, we compared the number of times that experts and novices noticed each of the 13 components of instruction (research question 1), provided evaluations of a lesson with reasoning (research question 2), and provided a suggestion about how to improve a lesson with reasoning (research question 2). We fit generalized linear models with one of these counts as the response variable and expertise (i.e., expert or novice) as the explanatory variable. We began by fitting a generalized linear model with Poisson-distributed errors and testing for goodness of fit with a chi-squared test. The null hypothesis of this test is that the model fits the data (UCLA: Statistical Consulting Group, n.d.). Therefore, we interpreted a $p < 0.05$ as indicative of poor model fit. A Poisson model assumes that mean and variance are equal. However, a variance larger than the mean is common in count data like ours and is referred to as overdispersion (Crawley, 2007). In cases in which a Poisson model was not a good fit for the data, we fit a negative binomial model and conducted another goodness-of-fit test. Negative binomial models fit an additional parameter, theta, to quantify overdispersion.

Making multiple comparisons to test a single hypothesis inflates the probability of false positives (i.e., type II error rate). Therefore, we corrected for multiple comparisons when necessary. We fit 13 models to answer research question 1,

TABLE 1. Example quotes demonstrating evaluations and suggestions with and without reasoning

Evaluations with reasoning	Evaluations without reasoning
"It appears the question builds upon a prior activity—therefore, students are being asked a question that is appropriate for their experience. It also appears that the questions serve multiple purposes—students are processing a concept they addressed in a previous course (checking their knowledge) AND they are being asked to 'reach' or extend that understanding a little beyond what was explicitly taught (analyzing new data in lane 6)."	"The instructor activated prior knowledge by asking students to recall their breakout experience. This is important for learning."
"The instructor was responsive to student questions and troubles, and took care to move around the classroom. That creates a more equitable class, in which students in the back still interact with the instructor. Clearly, the students are comfortable asking for help."	"The room didn't facilitate easy access by the instructor to the students. It would be easy for students in the middle of a row or in the back to get lost."
Suggestions with reasoning	**Suggestions without reasoning**
"Ask a different question in the whole group discussion that is more open-ended. Instead of "why is the answer 'no'?" and hearing from one male, I would ask "how would you explain to a colleague how you did this task?" and ask to hear from three different people. This would be an improvement for so many reasons, including having more students produce language to explain their answer, which often helps struggling students, and not portraying that the instructor is just looking for the right answer."	"Also, rather than listing the commonly believed difference between humans and other species on her slide, the professor could have asked the students to generate a list."
"The instructor asked a student to volunteer the answer, "Does anyone know?" Only a brave student who was confident he had the right answer would respond. It would have been more helpful for the teacher to have an iClicker question for all to answer, or for each student to write an answer on a piece of paper, then randomly choose one sheet to read in front of the class so as not to embarrass/praise any one student."	"Asking someone to volunteer answer. Instead do clicker or ask multiple answers before telling correct or incorrect."

and adjusted using the Holm (1979) method of correcting for multiple comparisons. We present back-transformed regression coefficients and 95% confidence intervals for these coefficients for each model we fit. In this context, back-transformed coefficients are multiplicative differences between experts and novices.

We also used quantitative analyses to compare expert and novice participants on characteristics relevant to our study, using Welch's two-sample t tests, Wilcoxon's rank-sum tests with continuity correction, and Fisher's exact tests. We compared demographics, course size, years of teaching experience, teaching professional development experience, and teaching practice as reported in section 3 of the TPI.

RESULTS

Our final sample included 14 experts and 29 novices and included only participants for whom complete data were available. Seventy-nine percent of these instructors identified as women, and this did not differ between experts and novices. Racial and ethnic diversity was limited in both groups. No participants identified as Hispanic and only five (12%) identified with any race besides white. Mean course size did not differ between experts and novices ($t = 1.37$, $p = 0.19$, M = 243, SD = 150). As expected given our inclusion criteria, experts reported that they had used active learning in large courses for significantly more years than novices (W = 400, $p < 0.0001$, Wilcoxon rank sum test). Experts reported using active-learning instruction for a median of 8 years (SD = 4.0), and novices reported a median of 2 years (SD = 1.2). Experts had also taught college biology courses for more terms (i.e., quarters or semesters) than novices (W = 322.5, $p = 0.002$). Experts reported teaching college biology courses for a median of 22 terms (SD = 18.1) compared with a median of 12 terms (SD = 12.4) for novices. Most participants (86%) had completed 40 or more hours of teaching professional development ($p = 0.65$), but experts were more likely to have led teaching professional development ($p < 0.0001$). Finally, experts reported using significantly more research-based practices in large undergraduate biology courses than did novices ($t = 3.5$, $p = 0.0012$). The mean score on Section 3 of the TPI was 12.5 (SD = 1.7) for experts and 10.1 (SD = 2.8) for novices, out of 15 possible points.

Research Question 1: How Do Experts and Novices Differ in *What* They Notice as They Analyze Active-Learning Lessons?

Experts noticed four components of instruction significantly more often than novices, including holding students accountable, planning for topic-specific difficulties, monitoring and responding to student thinking, and creating opportunities for generative work (Table 2). Our analyses estimate that novices paid more attention to some components than did experts, but these differences were not statistically significant. The 13 components of active-learning instruction that participants noticed are described and illustrated with quotes in Table 2. We provide additional detail about the four components that experts attended to significantly more often than novices. Finally, we describe the variation among experts in what they noticed as they analyzed active-learning lessons.

Holding Students Accountable. Participants attended to how instructors motivated students to participate and work during lessons by holding them accountable. They explained that students only benefit when they actually think about the questions posed to them. Participants noticed and suggested multiple approaches to holding students accountable for working during class time. This participant explained that students might be more motivated to work if they knew they would be turning in their work.

TABLE 2. Thirteen components of instruction noticed by active-learning instructors, with descriptions, illustrative quotes, and estimated differences in the frequency with which experts and novices noticed each component with associated 95% confidence intervals

Component	Participants discussed…	Example quote	Estimate (95% CI)[a]
Holding students accountable	instructor behaviors that impact students' motivation to engage and work in class by holding them accountable.	"Students were writing responses on a card, so I am assuming that the instructor is collecting these cards following the exercise. This creates accountability. The students know that the instructor might read how they answered the question, so they don't just check out during this activity."	5.8** (2.2–17.9)
Planning for topic-specific difficulties	difficulties students have in learning specific topics in biology, effective approaches for helping them overcome topic-specific difficulties, and grounding topics within specific biological contexts.	"I think using this activity was effective because it demonstrated for students in a practical way the effect of randomness on allele frequency changes; randomness is a very difficult concept for learners (ref., Klymkowsky) and seeing the process in action may, in fact, give students a better chance of grasping the concept."	4.7** (2.1–11.4)
Monitoring and responding to student thinking	whether and how instructors monitored student thinking while students were working, and if they used this knowledge to inform instructional decisions.	"I would have walked around the class to listen in or chat with groups as they discussed, then selected a few students to explain what they discussed. I believe this is a better way to get a sense of how the class is doing, rather than allowing a few students to volunteer answers."	4.4** (2.0–10.8)
Fostering community	how instructor behaviors and decisions can motivate students by making them feel more comfortable, feel a sense of belonging in the class, and feel that the instructor values them and their ideas.	"The instructor used student names, she asked for a volunteer who had not already shared, when talking in a small group she helped students use one another as resources, and checked in with students about how the activity was going re time. Community culture is an important part of being able to work hard and take chances in class, which is important for learning."	2.6 (1.1–6.2)
Building links between tasks	how the instructor helped students recognize links between tasks, and the value of making links explicit.	"The specific references to what the students had learned and how this activity would … stretch their knowledge was excellent."	2.1 (0.9–5.0)
Creating opportunities for generative work	the level of cognitive engagement of students, including when the instructor gave students responsibility for constructing their own ideas and engaging in scientific practices, either alone or with their peers.	"When a student identified a question they still had, the instructor affirmed that the student had identified the difficult question and asked her to talk to another student about their ideas. This, again, places the onus of learning on the student. If an answer comes out of an instructor's mouth, the student assumes it's correct and just writes it down. They've learned very little."	1.8* (1.3–2.4)
Making content relevant to students	whether content was likely to be interesting and relevant to students, thus motivating their participation.	"Content-wise, she set up an interesting conundrum-we are 'not special.' This acts as a hook that pulls people in."	1.7 (0.7–4.1)
Increasing equity	whether all students had the chance to participate in class by highlighting instructor behaviors that invited and allowed equitable engagement in individual work and whole-class discussions.	"One thing I would do differently would be to wait longer before calling on a student volunteer. Unless many hands went up immediately that were not in view, it seemed the instructor immediately called on the first volunteer to raise their hand. In my own experience, this can lead to the same small group of students dominating whole-class discussions."	1.3 (0.7–2.5)
Prompting metacognition	instruction that helps students recognize what they know and what they do not know, and provides guidance about how to monitor their own thinking and plan their learning.	"The instructor reviewed the basics, and then explicitly supported student metacognition by having them explain the results (stating that if they can't, they need to review). This helps students self-assess their progress, and models the type of behavior they should have throughout their courses."	1.4 (0.5–3.3)
Setting up lesson logistics	how the instructor laid out lesson expectations and instructions and managed time to keep students focused and not overloaded.	"The instructor also had a good set-up for the activity, clearly articulating the instructions and also the PURPOSE of the activity, which is helpful so that students have a clear goal in mind for why they are writing and talking with their neighbors."	1.1 (0.7–1.8)

Continued

TABLE 2. Continued

Component	Participants discussed…	Example quote	Estimate (95% CI)[a]
Creating opportunities for active work	whether students were physically doing something during a lesson, such as an "activity."	"Instead of a lecture, he chose to illustrate a complex and not intuitive concept using a hands-on activity. This was meant to engage the students with the material, at least as I understand it."	0.8 (0.5–1.2)
Monitoring lesson logistics	the instructor circulating through the classroom to determine how much time students need and to respond to confusions about what students are supposed to be doing.	"The instructor is perceptive to the amount of time students need to work on the task as is evidenced by asking them if they have had enough time. This will prevent students from rushing or waiting too long."	0.7 (0.3–1.4)
Materials and delivery	surface features of the classroom, including the materials, instructor's delivery when speaking, equipment, and physical space.	"The text on the board might be hard for students to see in the back. This can be improved by using a different color or writing that text on the slide or document viewer."	0.6 (0.2–1.2)

[a]Estimates are interpreted as follows: experts noticed how instructors held students accountable 5.8 times more often than did novices.
*$p < 0.05$, adjusted for multiple comparisons.
**$p < 0.01$, adjusted for multiple comparisons.

> It was good that the instructor made the students commit to their answer on paper. That this would be taken up, making each student feel responsible, was hinted at by the fact that they were instructed to write their ID on the paper.

Another participant explained that associating course credit with in-class work can motivate students to participate, even if they do not yet see the value of the work.

> I use a lot of strategies in my introductory course to "force" them to participate, in the hopes that they will soon realize how much this can help their learning process. I have heard other faculty discount this as "offering points for everything," but I have seen these small incentives really change student's behavior early in their educational career.

Participants also explained that the instructor moving around the room while students worked can encourage participation. Some noticed an instructor posing questions to students while circulating, making students feel obligated to be prepared to discuss their thinking with the instructor. Others noted that less engaged students could be brought back into the discussion if the instructor talked with them. For instance, one observed,

> While a few students in the front of the frame were clearly discussing the question, the students behind them were not engaged. The girl threw up her hands indicating she didn't know and then played with her hair. The instructor could have circulated among the students or targeted prompting, helpful questions to students who were not engaged in the activity.

A few participants noted that calling only on volunteers for discussions does not hold all students accountable for working, but randomly calling on students can accomplish this.

> Most importantly, engage ALL students in the decision-making process. It appears that the eager students up front were involved but the rest were waiting until someone else did the explaining. This could be done using personal response systems, or by just having everyone write down their answer and calling RANDOMLY on the class. Thus, they know that they COULD be called upon.

Planning for Topic-Specific Difficulties. Participants noticed events related to teaching and learning that went beyond disciplinary knowledge or pedagogical knowledge. Rather, participants displayed knowledge about teaching- and learning-specific concepts (i.e., topics) within biology. This knowledge stood out, because much of what participants discussed was more general pedagogical knowledge.

Participants paid attention to how instructors taught specific topics, discussing difficulties students commonly have when learning a topic and effective approaches for helping students overcome those difficulties. The topics covered in the three videos that participants analyzed were evolutionary relationships between humans and other great apes, genetic drift, and Golgi structure and function. Most of these comments concerned teaching and learning genetic drift, including this one, which discusses both student difficulties and effective teaching strategies:

> Instructor attempted to use an in-class activity to teach students about genetic drift. Of all of the mechanisms of evolution, this is the most difficult one for students to understand and I have found that activities in which students can see changes in allele frequencies developing, without natural selection or gene flow, are really powerful for allowing students to develop an intuitive sense of this concept.

A few participants discussed how to improve instruction about the phylogenetic relationships between humans and other great apes, including this one, which suggests some changes to help students overcome an inaccurate idea:

> Her explanation of what it would look like on a phylogeny if humans evolved from chimps was technically accurate, but I don't know if it would have been helpful for a student that

> actually thought that humans evolved from chimps because she said it pretty quickly. I would have a visual ready on the next slide that illustrated what a phylogeny would look like if humans evolved from chimps. Then I could compare it to the actual phylogeny to highlight the important differences.

Participants also considered the degree to which problems posed to students were sufficiently grounded in biological contexts. Most of these comments focused on one specific lesson. In this lesson, the instructor introduced an activity that required students to roll a die and graph allelic frequency changes based on the numbers they rolled. The activity was intended to help students understand genetic drift, but participants noted that it was not sufficiently grounded in a biological context to allow students to fully grasp the abstract concept being practiced in the activity and relate it biology.

> It appeared that the class started without much intro to discuss genetic drift or how the activity linked to it. In this case, most of the discussion was saved for after the activity—the instructor made several comments about moving on and coming back to concepts later. Students regularly complete activities without understanding how those activities link to course concepts. It sounds like student comments in this video support this. I believe that an introductory review of genetic drift (since it appears they have already covered it), probability, and randomness with respect to phenotype (preferably one where students provided most of the information) would help link the activity to concepts better.

Monitoring and Responding to Student Thinking. Participants noticed whether instructors took steps to learn about student thinking during a lesson and, ideally, to adapt instruction based on what they learned. Participants explained that it was important for instructors to be aware of the thinking of all students. Some participants noted that they could not analyze the effectiveness of a lesson without knowing what students were thinking. They discussed several approaches to reveal student thinking, including polling using clickers or raised hands, collecting student work, and circulating to talk to students as they worked. For example,

> I might have started the discussion by posing the question "did we evolve from chimps" and polling the audience about what they thought in order to first assess what students understood already.

One participant explained how hearing from multiple students instead of just one student gives a clearer picture of students' current thinking:

> I would have walked around the class to listen in or chat with groups as they discussed, then selected a few students to explain what they discussed. I believe this is a better way to get a sense of how the class is doing, rather than allowing a few students to volunteer answers … the same few students tend to do so every time, and other groups may not be staying on task.

A handful of participants went a step further and addressed how an instructor could use what he or she learned to inform instruction in real time and in the future. For example,

> I would maybe have had a "clicker question" at the ready to present to the class after they discussed among themselves. The answers to the clicker question could have had two reasonable explanations and three unreasonable explanations. After seeing how the class votes, you can determine whether they sufficiently understand and you can move on, or you can identify specific misunderstandings and address them.

Creating Opportunities for Generative Work. Many participants considered whether students were asked to generate ideas that went beyond those provided by the instructor (i.e., generative cognitive engagement). They attended to opportunities students had to do generative work during each lesson, including the problems posed to students and how the instructor facilitated student work during class time. Participants praised instructors for asking students to figure out why an answer was correct rather than just asking them to recognize a correct answer.

> The most effective learning experience came when the instructor prompted students to work with their neighbor to come up with an explanation for a problem. This was good for a number of reasons. First, the emphasis was on providing reasoning rather than coming up with the correct answer. In fact, the answer was given up front by the teacher and the teacher then prompted students specifically to talk through the reasoning for the answer with a neighbor.

One type of generative work that participants noticed were opportunities students had to engage in scientific practices. They commented on instruction that required students to practice analyzing data, interpreting figures, evaluating study design, and understanding the purpose of controls in an experiment. Most participants did not provide a rationale for these statements. However, some participants valued instruction that gave students the chance to engage in scientific practices because they saw this as a key goal of undergraduate biology instruction. These quotes emphasized that students had the opportunity to engage in the real work of scientists, leading to richer understanding of the process and nature of science.

> The activity that the instructor is having the students engage in is a higher order Bloom's problem. They are being asked to use their knowledge to analyze real data. They are being asked to practice being real scientists. This should be the goal of all biology classes, not just upper division ones such as this, but especially in upper division classes, the majority of class time should be spent on higher order Bloom's problems and not on regurgitation of facts.

In addition to attending to the tasks instructors asked students to complete, participants noticed how instructor decisions prompted students to continue to do generative work throughout class time. For example, an instructor can respond to students' questions without providing an answer. This approach keeps the onus of learning on the student, rather than transferring the intellectual work back to the instructor. The instructor may pose follow-up questions rather than supplying the correct answer or can guide a student to work with a peer to answer the question. Another instructor technique that keeps

the responsibility for intellectual work with the students is asking them to lead the wrap-up of an active-learning exercise, rather than the instructor providing the reasoning and conclusion. None of the videos showed this, so these comments discussed a missed opportunity. One participant said,

> In explaining the rationale, the instructor again took over on the explanation instead of relying on students to fill in the gaps. Several students could have constructed a complete response for the class instead.

Variation among Experts. In addition to revealing differences between experts and novices, our approach uncovered variation among experts. Experts varied in the number of components of instruction they noticed as they analyzed lessons. Seven (50%) experts addressed nine or more components of instruction in their analyses, but some addressed just four or five. Among those experts who noticed fewer components, they mostly commonly did not pay attention to monitoring lesson logistics, holding students accountable, prompting metacognition, making content relevant to students, and materials and delivery.

Research Question 2: How Do Expert and Novice Active-Learning Instructors Differ in Their Ability to *Use Reasoning* to Support Their Evaluations of and Suggestions for Improving Active-Learning Lessons?

Experts and novices differed in the degree to which they were able to support what they noticed with reasoning. Experts made more evaluations with reasoning and provided more suggestions for improving active-learning lessons that were supported by reasoning (Table 3).

One potential explanation for differences between experts and novices is that experts wrote more than novices in response to questions in the lesson-analysis survey, providing greater insight into their knowledge. We tested this alternative explanation by comparing the number of words used to respond to all questions on the lesson-analysis survey. The number of words used by experts and novices did not differ ($t = 1.15$, $p = 0.26$, M = 597, SD = 390).

DISCUSSION

Our findings provide the first empirical insight about what teacher knowledge is associated with effective active-learning instruction in large undergraduate STEM courses. We hold the perspective that college instructors are learners when it comes to becoming effective active-learning instructors, and this work pinpoints the most important learning objectives for teaching professional development that aims to support active-learning instruction. We did not discover any ideas that were used significantly more frequently by novices than experts, suggesting that what distinguishes novices is a lack of knowledge, rather than particular unproductive ideas. This work contributes foundational knowledge about knowledge for active-learning instruction in undergraduate STEM and can ground the design and testing of evidence-based preparation and support for instructors. We discuss our findings in relation to prior work to provide additional context and theoretical grounding for interpretation and future direction. Specifically, we address the role of student motivation, student thinking, and generative work in large active-learning classrooms. We end the paper by proposing avenues for future work, considering limitations, and making concluding remarks.

TABLE 3. Estimated differences between experts and novices in the frequency with which they provided reasoning and associated 95% confidence intervals

	Estimate[a] (95% CI)
Evaluation with reasoning	2.9*** (1.6–5.5)
Suggestion with reasoning	3.8** (1.6–9.6)

[a]Estimates are interpreted as: experts provided 2.9 times as many evaluations with reasoning as novices.
**$p < 0.01$.
***$p < 0.001$.

Student Motivation in Large Active-Learning Courses

Our findings indicate that expert active-learning instructors were, on average, more aware than novices of some best practices related to student motivation. The importance of attending to student motivation in large active-learning courses has been recognized previously. Eddy *et al*. (2015) reviewed investigations of the relationship between student outcomes and specific features of active learning in large undergraduate STEM courses. Student motivation underlies two of the four dimensions of best practices they identified as important to effectiveness: accountability and reducing apprehension (Eddy *et al.*, 2015). These dimensions have similarities to what our participants discussed about holding students accountable and fostering community, respectively. Experts noticed whether instructors held students accountable significantly more often than novices did, but more than half of experts did not discuss accountability at all. Together, our work and prior work suggest that attending to student motivation is likely important to effective active-learning instruction and that this may be an important learning objective for novices and some experts.

Some philosophical beliefs about the role and responsibilities of college instructors may act as a barrier to adopting practices that promote student motivation to work during class time. One of our participants described colleagues who oppose allowing students to earn points for answering in-class questions, because these colleagues see it as "offering points for everything." Empirical work investigating the details of these ideas among college instructors and the impact of such ideas on the adoption of active learning and other evidence-based strategies is needed. While some ideas about teaching and learning are strongly held and unlikely to change, other ideas may be better targets for change.

Student Thinking in Large Active-Learning Courses

More effective instructors better recognize the prominent role that student thinking plays in effective active-learning classrooms. Two of the components of instruction that experts attended to more frequently than novices deal with student thinking: monitoring and responding to student thinking and planning for topic-specific difficulties (Table 2). Many studies of teacher noticing among K–12 instructors demonstrate that teachers turn their attention to student thinking and the

relationship between student thinking and teaching strategies as they gain experience and expertise (e.g., Sherin and van Es, 2008; van Es, 2011; Kisa and Stein, 2015). Attending to student thinking not only allows an instructor to make decisions about how to proceed with a lesson, it allows the instructor to provide students with immediate feedback about their thinking. These two ideas are encompassed by the term "formative assessment." Formative assessment refers to diverse methods used to gather information to be used as feedback. This feedback allows students and instructors to modify teaching and learning activities (e.g., Black and Wiliam, 2006; Offerdahl and Montplaisir, 2013). Our findings suggest expert active-learning instructors think more about using formative assessment to monitor and respond to student thinking.

More effective active-learning instructors may also be better equipped to anticipate difficulties students are likely to encounter when learning a particular concept and to address these difficulties in real time. The component we have called "planning for topic-specific difficulties" relates to a theoretical construct of teacher knowledge called pedagogical content knowledge (PCK). PCK is knowledge at the intersection of disciplinary content knowledge and pedagogical knowledge. The researcher who originally proposed PCK described it as "the category most likely to distinguish the understanding of the content specialist from the pedagogue" (Shulman, 1987, p. 8). In other words, PCK is unlikely to be developed through training in the discipline. It is constructed through preparation for teaching and reflective teaching experiences (e.g., van Driel *et al.*, 1998; Chan and Yung, 2015). The majority of research on teacher knowledge in recent decades has focused on PCK.

PCK is topic-specific knowledge of teaching and learning, meaning that instructors may need distinct PCK for each topic they teach (e.g., natural selection, speciation, genetic drift; Gess-Newsome, 2015). Researchers have outlined the components of PCK in various ways (e.g., Shulman, 1987; Magnusson *et al.*, 1999; Park and Oliver, 2008) but widely agree that PCK includes knowledge of 1) student difficulties in learning a particular topic and 2) instructional strategies and representation for teaching a topic (e.g., Magnusson *et al.*, 1999; Park and Oliver, 2008; Alonzo and Kim, 2015; Chan and Yung, 2015). There is extensive empirical support that PCK influences instruction (e.g., Park *et al.*, 2011) and is associated with student learning (e.g., Hill *et al.*, 2005; Sadler *et al.*, 2013; Blömeke *et al.*, 2015), but most of this work has focused on K–12 teachers.

PCK has been investigated much less often among college instructors, but a few studies provide valuable insights into what knowledge may be important for effective active-learning instruction. In-depth, semester-long studies of college mathematics instructors adopting inquiry-based curricula for the first time found that instructors struggled because they lacked awareness of likely student difficulties with specific topics, they faltered in trying to make sense of students' ill-formed reasoning in-the-moment while facilitating discussions, and they could not recognize or figure out how the ideas that students contributed could be relevant to lesson goals (Wagner *et al.*, 2007; Speer and Wagner, 2009; Johnson and Larsen, 2012). Together with our work, this is compelling evidence that knowing the difficulties students are likely to encounter with specific topics is important for planning and enacting effective active-learning instruction in college courses. This has implications for teaching professional development for college STEM faculty. If some teacher knowledge critical to effective instruction is topic specific, generalized teaching professional development programs provided by centers of teaching and learning may be insufficient alone. Instructors may need formal preparation for teaching in the discipline or even for teaching in their area of specialization (e.g., evolutionary biology, ecology, cellular biology).

Generative Work in Large Active-Learning Courses

Expert active-learning instructors paid more attention to whether instruction provided opportunities for students to construct their own ideas (e.g., generate explanations with reasoning, apply knowledge to a new scenario, analyze and interpret data). This finding aligns with the ICAP framework, which posits that students learn more from generative work than from physically active work (Chi and Wylie, 2014). The estimated differences between experts and novices were not as large as for other distinguishing components, and 93% of novices mentioned generative work in their analyses at least once, suggesting that novices may begin to appreciate this fundamental idea about active-learning instruction before developing other important knowledge. Nonetheless, teaching professional development for college STEM faculty should constantly emphasize the learning benefits of tasks that require students to generate their own ideas, rather than simply fostering activity, until this is common knowledge across the academy.

Future Work: Fostering the Development of Knowledge for Active-Learning Instruction

This work identifies differences between experts and novices, but not how novices come to be experts. The components of instruction noticed more frequently by experts than novices can be considered learning objectives for novices, but future work must investigate how novices develop this knowledge and what interventions can facilitate knowledge development. Though experience is undoubtedly critical, teacher preparation may also play an integral role in accelerating expertise development. The undergraduate STEM education community will benefit from drawing on the knowledge base that has resulted from decades of research on the preparation and development of K–12 instructors. For example, teacher noticing has been used as a theoretical framework for the design of preservice and in-service teacher training. Teaching professional development in which instructors learn to critically analyze videos of lessons has been effective at fostering teacher knowledge and promoting student-centered instruction among K–12 teachers (van Es and Sherin, 2008; Sherin and van Es, 2008; Zhang *et al.*, 2011). One program, which taught teachers to ground their analyses in learning objectives and student progress toward those objectives, produced teachers who were more skilled at eliciting student thinking and letting student thinking guide instruction, even very early in their teaching careers (Santagata and Yeh, 2014). This model, and others, could be adapted for college STEM instructors, with the aim of fostering the construction of knowledge that is associated with effective active-learning instruction.

Limitations

This work has several limitations that are important to consider when interpreting the results. First, a lesson-analysis approach

to studying teacher knowledge reveals knowledge that instructors possess and can use to evaluate another instructor's teaching, but it does not reveal what knowledge they actually apply to their own instruction. Nevertheless, prior studies have found positive relationships between evaluating videos of other instructor's lessons and student learning in a teacher's own course (e.g., Kersting *et al.*, 2012). Additionally, the ability of experts to provide reasoning for what they noticed indicates they not only possess knowledge of what to do or how to do it, but also of why it works. Rogers (2003), who studies the adoption of innovations across diverse contexts, refers to this as "principles knowledge" and contends that it may be important to adapting innovations and persisting in using them. Future studies of teacher knowledge for active-learning instruction in undergraduate STEM can build on this work by directly examining teacher knowledge in action.

A second limitation is the number of expert active-learning instructors we were able to identify and recruit to participate. We opted to maintain strict criteria for experts at the expense of a larger sample. We considered this important, because it increases the trustworthiness of our findings about what knowledge is associated with expertise. The downside of this choice was reduced statistical power to detect differences between experts and novices. Remedying this problem in future work will likely require researchers to systematically measure student learning gains in participants' courses themselves.

A third limitation relates to our expert–novice approach. Expert–novice studies are grounded in stage theory, which posits that individuals move through a pattern of distinct stages over time as they develop professional expertise (Dreyfus and Dreyfus, 1986). Such studies have been critiqued for omitting individuals in intermediate levels of development and for assuming that all individuals develop in the same direction and through the same stages (Engeström *et al.*, 1999; Dall'Alba and Sandberg, 2006). We found it necessary to start by investigating experts and novices, because we could set and apply consistent criteria to distinguish these groups from each other. This approach allowed us to identify knowledge that is associated with mastery. Future work must investigate how this knowledge develops over time. Future researchers should not assume that all active-learning instructors develop along the same trajectory. Experts varied in the knowledge they brought to their lesson analyses, which could indicate that experts also have more room for growth. It could also mean that not all active-learning instructors need the same knowledge to be effective. Future work will be necessary to make that distinction and to reveal the varied trajectories of expertise development.

A fourth potential limitation of this approach to quantitatively comparing experts and novices is that the ability to provide reasoning—rather than simply to notice—is a hallmark of expertise and may be more likely to be related to teaching practices. In addition to determining how often each participant noticed each of the 13 components of instruction, we counted how many times each participant noticed *and* reasoned about each of the 13 components. This results in many more zeros in the data set, because most participants notice some components without providing reasoning. We fit 13 generalized linear models using the same approach described in the data analysis, but with counts of noticing with reasoning as the response variables. Two components of instruction were significantly different between experts and novices: planning for topic-specific difficulties and holding students accountable. Comparing how often experts and novices noticed and reasoned about monitoring and responding to student thinking and building links between tasks both resulted in $p < 0.10$. As described in our first limitation, an important next step in this research will be examining the knowledge instructors use in their own teaching.

We focused on large courses in this study, because they are particularly challenging contexts and contexts in which instructors may be especially reluctant to try active learning. We cannot make claims about the kinds of teacher knowledge that are most important in smaller courses.

Finally, we caution readers about generalizing our findings regarding novices to all college biology faculty. All of the instructors in this study had made a decision to try active-learning instruction. Furthermore, 83% of novices had engaged in 40 or more hours of teaching professional development, which may not be representative of typical college faculty. We might find larger differences in teacher noticing if we compared expert active-learning instructors and a random sample of college instructors.

CONCLUSIONS

The potential positive impact of active-learning instruction on student learning and retention in undergraduate STEM has *not* been widely realized. High-profile calls for incorporating active-learning instruction have not been accompanied by high-profile calls to fundamentally transform how we prepare and support college STEM instructors. Yet most college instructors have little or no preparation in teaching. Furthermore, active-learning instruction is fundamentally different from the traditional lecture approach most college instructors experienced as students. Achieving the gains active-learning instruction promises for increasing the diversity and preparation of STEM undergraduates will likely require widespread and systemic reform so that all college instructors have the opportunity *and* imperative to develop deep knowledge of how people learn. The knowledge development of college instructors, like that of any learner, will be facilitated by evidence-based learning opportunities. This work is a small first step in discovering what teacher knowledge is critical to effective active-learning instruction in undergraduate STEM. Future work must build on this and begin exploring how we can support instructors in developing this critical knowledge.

ACKNOWLEDGMENTS

Thank you to our research participants, including those who were part of pilot studies, without whom this work would not be possible. We also owe a debt of gratitude to the instructors who were willing to have their classroom filmed and shown to other college instructors. Thank you to the University of Georgia BERG group, especially Julie Stanton and Jennifer Thompson, for their continuous feedback and support. Thanks also to our lab colleagues for inspiration, feedback, and support. Thank you to undergraduate assistants who contributed to this project, including Logan Gin, Aarushi Lal, Adebimpe Atanda, and Andrew Potocko. Thank you to the monitoring editor and anonymous reviewers for insightful feedback that helped us improve

our manuscript. Partial support for this work was provided by the National Science Foundation's Improving Undergraduate STEM (IUSE) program under award No. 1504904. 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 National Science Foundation. Partial funding was provided by the University of Georgia Office of STEM Education mini-grant.

REFERENCES

Alonzo, A. C., & Kim, J. (2015). Declarative and dynamic pedagogical content knowledge as elicited through two video-based interview methods. *Journal of Research in Science Teaching*, *53*(8), 1259–1286.

Ambrose, S. A., Bridges, M. W., DiPietro, M., & Lovett, M. C. (2010). *How learning works: Seven research-based principles for smart teaching*. San Francisco: Jossey Bass.

American Association for the Advancement of Science. (2011). *Vision and change in undergraduate biology education: A call to action*. Washington, DC.

Anderson, D. L., Fisher, K. M., & Norman, G. J. (2002). Development and evaluation of the conceptual inventory of natural selection. *Journal of Research in Science Teaching*, *39*(10), 952–978.

Andrews, T. C., Conaway, E. P., Zhao, J., & Dolan, E. L. (2016). Colleagues as change agents: How department networks and opinion leaders influence teaching at a single research university. *CBE—Life Sciences Education*, *15*(2), ar15.

Andrews, T. M., Leonard, M. J., Colgrove, C. A., & Kalinowski, S. T. (2011). Active learning not associated with student learning in a random sample of college biology courses. *CBE—Life Sciences Education*, *10*(4), 394–405.

Black, P., & Wiliam, D. (2006). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5,* 7–74.

Blömeke, S., Hoth, J., Döhrmann, M., Busse, A., Kaiser, G., & König, J. (2015). Teacher change during induction: Development of beginning primary teachers' knowledge, beliefs, and performance. *International Journal of Science and Mathematics Education*, *13,* 287–308.

Chan, K. K. H., & Yung, B. H. W. (2015). On-site pedagogical content knowledge development. *International Journal of Science Education*, *37*(8), 1246–1278.

Chi, M. T. H. (2009). Active-Constructive-Interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, *1,* 73–105.

Chi, M. T. H., Kang, S., & Yaghmourian, D. L. (2017). Why students learn more from dialogue- than monologue-videos: Analysis of peer interactions. *Journal of the Learning Sciences*, *26*(1), 10–50.

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219–243.

Crawley, M. J. (2007). *The R book*. Chichester, UK: Wiley.

Dall'Alba, G., & Sandberg, J. (2006). Unveiling professional development: A critical review of stage models. *Review of Educational Research*, *76*(3), 383–412.

Dancy, M., Henderson, C., & Turpen, C. (2016). How faculty learn about and implement research-based instructional strategies: The case of peer instruction. *Physical Review*, *12,* 010110.

Dreyfus, H., & Dreyfus, S. (1986). *Mind over machine: The power of human intuitive expertise in the era of the computer*. New York: Free Press.

Dunlap, W. P., Cortina, J. M., Vaslo, J. B., & Birke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures design. *Psychological Methods*, *1*(2), 170–177.

Eddy, S. L., Converse, M., & Wenderoth, M. P. (2015). PORTAAL: A classroom observation tool assessing evidence-based teaching practices for active learning in large science, technology, engineering, and mathematics classes. *CBE—Life Sciences Education*, *14,* ar23.

Engeström, Y., Miettinen, R., & Punamäki, R. L. (Eds.) (1999). *Perspectives on activity theory*. Cambridge University Press.

Gess-Newsome, J. (2015). A model for teaching professional knowledge and skill including PCK: Results of the thinking from the PCK Summit. In Berry, A., Friedrichsen, P., & Loughran, J. (Eds.), *Re-examining pedagogical content knowledge in science education* (pp. 28–42). New York: Routledge.

Glynn, S. M., Taasoobshirazi, G., & Brickman, P. (2009). Science motivation questionnaire construct validation with nonscience majors. *Journal of Research in Science Teaching*, *46*(2), 127–146.

Henderson, C., & Dancy, M. H. (2009). The impact of physics education research on the teaching of introductory quantitative physics. *AIP Conference Proceedings*, *1179,* 165–168. doi: 10.1063/1.3266705

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, *42*(2), 371–406.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6,* 65–70.

Johnson, E. M. S., & Larsen, S. P. (2012). Teacher listening: The role of knowledge of content and students. *Journal of Mathematical Behavior*, *31,* 117–129.

Kane, R., Sandretto, S., & Heath, C. (2004). An investigation into excellent tertiary teaching: Emphasising reflective practice. *Higher Education*, *47*(3), 283–310.

Kersting, N. B. (2008). Using video clips of mathematics classroom instruction as item prompts to measure teachers' knowledge of teaching mathematics. *Educational and Psychological Measurement*, *68*(5), 845–861.

Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R., & Stigler, J. (2012). Measuring usable knowledge: Teachers analyses of mathematics classroom videos predicts teaching quality and student learning. *American Educational Research Journal*, *49*(3), 568–589.

Kisa, M. T., & Stein, M. K. (2015). Learning to see teaching in new ways: A foundation for maintaining cognitive demand. *American Educational Research Journal*, *52*(1), 105–136.

Magnusson, S., Krajcik, J., & Borko, H. (1999). Nature, sources, and development of pedagogical content knowledge for science teaching. *Examining pedagogical content knowledge* (pp. 95–132). Dordrecht, Netherlands: Springer.

McAlpine, L., & Weston, C. (2000). Reflection: Issues related to improving professors' teaching and students' learning. *Instructional Science*, *28,* 363–385.

McAlpine, L., Weston, C., Berthiaume, D., & Fairbank-Roch, G. (2006). How do instructors explain their thinking when planning and teaching? *Higher Education*, *51,* 125–155.

Menekse, M., Stump, G., Krause, S., & Chi, M. T. H. (2013). Differentiated overt learning activities for effective instruction in engineering classrooms. *Journal of Engineering Education*, *102*, 346–374.

Middlemis, M. J., Markey, J. C., & Ebert May, D. (2013). The other half of the story: Effect size analysis in quantitative research. *CBE—Life Sciences Education*, *12,* 345–351.

National Research Council. (2012). *Discipline-based education research: Understanding and improving learning in undergraduate science and engineering*. Washington, DC: National Academies Press.

Offerdahl, E. G., & Montplaisir, L. (2013). Student-generated reading questions: Diagnosing student thinking with diverse formative assessment. *Biochemistry and Molecular Biology Education*, *42*(1), 29–38.

Palmer, D. J., Stough, L. M., Burdenski, T. K. Jr., & Gonzales, M. (2005). Identifying teacher expertise: An examination of researchers' decision making. *Educational Psychologist*, *40*(1), 13–25.

Park, S., Jang, J-Y., Chen, Y-C., & Jung, J. (2011). Is pedagogical content knowledge (PCK) necessary for reformed science teaching? Evidence from an empirical study. *Research on Science Education*, *41,* 245–260.

Park, S., & Oliver, J. S. (2008). Revisiting the conceptualisation of pedagogical content knowledge (PCK): PCK as a conceptual tool to understand teachers as professionals. *Research in Science Education*, *38*(3), 261–284.

Pollock, S. J., & Finkelstein, N. D. (2008). Sustaining educational reforms in introductory physics. *Physical Review*, *4,* 010110.

Rogers, E. (2003). *Diffusion of innovations*, 5th ed.. New York: Free Press.

Sadler, P. H., Sonnert, G., Coyle, H. P., Cook-Smith, N., & Miller, J. L. (2013). The influence of teachers' knowledge on student learning in middle school physical science classrooms. *American Education Research Journal*, *50*(5), 1020–1049.

Santagata, R., & Angelici, G. (2010). Studying the impact of the lesson analysis framework on preservice teachers' abilities to reflect on videos of classroom teaching. *Journal of Teacher Education*, *61*(4), 339–349.

Santagata, R., & Yeh, C. (2014). Learning to teach mathematics and to analyze teaching effectiveness: Evidence from a video- and practice-based approach. *Journal of Mathematics Teacher Education, 17*(6), 491–514.

Schön, D. A. (1987). *Educating the reflective practitioner: Toward a new design for teaching and learning in the professions*. San Francisco: Jossey-Bass.

Sherin, M. G., Jacobs, V. R., & Phillip, R. A. (2011). Situating the study of teacher noticing. In Sherin, M. G., Jacobs, V. R., & Phillip, R. A. (Eds.), *Mathematics teacher noticing: Seeing through teachers' eyes* (pp. 3–13). New York: Routledge.

Sherin, M. G., & van Es, E. A. (2008). Effects of video club participation on teachers' professional vision. *Journal of Teacher Education*, *60*(1), 20–37.

Shi, J., Wood, W. B., Martin, J. M., Guild, N. A., Vicens, Q., & Knight, J. K. (2010). A diagnostic assessment for introductory molecular and cellular biology. *CBE—Life Sciences Education*, *9*(4), 453–461.

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, *57*(1), 1–23.

Smith, M. K., Vinson, E. L., Smith, J. A., Lewin, J. D., & Stetzer, M. R. (2014). A campus-wide study of STEM courses: New perspectives on teaching practices and perceptions. *CBE—Life Sciences Education*, *13*(4), 624–635. doi: 10.1187/cbe.14-06-0108

Smith, M. K., Wood, W. B., & Knight, J. K. (2008). The Genetics Concepts Assessment: A new concept inventory for gauging student understanding of genetics. *CBE—Life Sciences Education*, *7*(4), 422–430.

Speer, N. M., & Wagner, J. F. (2009). Knowledge needed by a teacher to provide analytic scaffolding during undergraduate mathematics classroom discussions. *Journal for Research in Mathematics Education*, *40*(5), 530–562.

Stains, M., & Vickrey, T. (2017). Fidelity of implementation: An overlooked yet critical construct to establish effectiveness of evidence-based instructional practices. *CBE—Life Sciences Education*, *16*(1), rm1.

Talanquer, V., Tomanek, D., & Novodvorsky, I. (2012). Assessing students' understanding of inquiry: What do prospective science teachers notice?. *Journal of Research in Science Teaching*, *50*(2), 189–208.

Turpen, C., & Finkelstein, N. D. (2009). Not all interactive engagement is the same: Variations in physics professors' implementation of peer instruction. *Physical Review*, *5,* 020101.

UCLA: Statistical Consulting Group. (n.d.). Poisson Regression: R Data Analysis Examples. Retrieved July 25, 2017, from https://stats.idre.ucla.edu/r/dae/poisson-regression/

van Driel, J. H., Verloop, N., & de Vos, W. (1998). Developing science teachers' pedagogical content knowledge. *Journal of Research in Science Teaching*, *35*(6), 673–695.

van Es, E. A. (2011). A framework for learning to notice student thinking. In Sherin, M. G., Jacobs, V. R., & Phillipp, R. A. (Eds.), *Mathematics teacher noticing: Seeing through teachers' eyes* (pp. 134–151). New York: Routledge.

van Es, E. A., & Sherin, M. G. (2008). Mathematics teachers' "learning to notice" in the context of a video club. *Teaching and Teacher Education*, *24,* 244–276.

Vermunt, J. D. (1996). Metacognitive, cognitive and affective aspects of learning styles and strategies: A phenomenographic analysis. *Higher Education*, *31,* 25–50.

Wagner, J. F., Speer, N. M., & Rossa, B. (2007). Beyond mathematical content knowledge: A mathematician's knowledge needed for teaching an inquiry-oriented differential equations course. *Journal of Mathematical Behavior*, *26,* 247–266.

Waldrop, M. M. (2015, July 15). Why we are teaching science wrong, and how to make it right. *Nature News Feature*.

Wieman, C., & Gilbert, S. (2014). The teaching practices inventory: A new tool for characterizing college and university teaching in mathematics and science. *CBE—Life Sciences Education*, *13*(3), 552–569. doi: 10.1187/cbe.14-02-0023

Zhang, M., Lundeberg, M., Koehler, M. J., & Eberhardt, J. (2011). Understanding affordances and challenges of three types of video for teacher professional development. *Teaching and Teacher Education*, *27*, 454–462.

Supplemental Material

CBE—Life Sciences Education

Auerbach *et al.*

APPENDIX A

Interview protocol for screening interviews

I will start by telling you a little bit about our study. Although active learning can lead to greater student learning than lecture alone, these strategies are not always effective. We want to dig a little bit more into this, so we are talking to instructors that use active learning to see how long they have been using these strategies, what these strategies look like in their classroom, and how they know when these strategies are working. We believe there is a difference in the knowledge used to implement active learning that results in the different outcomes. We hope to elicit this difference using a video-survey and apply our findings to designing professional development that is aimed at providing instructors with the knowledge and skills they need to become even more effective.

We are using these phone conversations to screen possible participants before referring them to the video survey. However, we have found that there is rich data given related to the screening process and would like your permission to record and use the responses for qualitative analysis. We hope to add another dimension to the knowledge instructors use by looking at the way they describe their own classrooms and instructional decision-making. We will only use information gathered during this phone call should you be deemed eligible and willing to participate in the video-survey. If you are deemed ineligible or choose not to participate in the video-survey, your records will be permanently deleted.

1. Do you teach any large enrollment biology courses?
2. How long have you been using active learning in those courses?
3. What types of active learning do you use, what does AL look like in your classroom?
4. How do you decide what to use?
5. How do you know when something is or isn't working?
6. Are there other ways you use feedback from students to inform your teaching?

APPENDIX B

Developing and Refining Lesson-Analysis Survey

The lesson-analysis survey created for this study included three videos of real active-learning lessons in a college biology course. Each video was accompanied by writing prompts to elicit teacher noticing. Prior lesson-analysis instruments guided our instrument development. We revised and refined the videos and writing prompts using both pilot data and expert feedback. We describe this process in detail below.

Video production

Videos for this instrument needed to be short to minimize participant burden, but also needed to show a section of a lesson that could stand alone. We also required high-quality audio and video and footage of both the instructor and students. We filmed numerous classrooms to collect sufficient and appropriate footage for our research goals. We used two or three cameras each time we filmed, allowing us to simultaneously capture what the instructor and students were doing. We also recorded audio in the classrooms using two to four microphones, including a lapel microphone to clearly capture the voice of the instructor, two shotgun boom microphones pointed at the students to capture student voices, and the microphone on one or more cameras to capture the collective sound of the classroom. We told students in the filmed courses that the videos would primarily be viewed by college instructors, and that designated parts of the room would not be filmed if they did not wish to appear in video. We produced over eight hours of classroom footage from which we created short (i.e., 3-5 minute) clips for the the survey.

We selected and edited video and audio footage to create clips that showed a variety of teaching decisions and practices. We began this process by identifying sections of footage that would be meaningful on their own and that included multiple practices, such as lecturing, small-group discussion, and whole-class discussion. We then spliced instructor footage with footage of students so that viewers had the opportunity to notice both. We added subtitles as necessary so the viewer always had access to what the instructor and featured students were saying. We also began each video with information about the class to provide context for viewers, including the institution type, course title, class size, and current topic. We used Adobe Premier Pro CC v9.2 for all video and audio editing.

Testing and refining lesson-analysis survey

We paired video clips of active-learning classrooms with prompts to elicit teacher knowledge. Prompts asked respondents to critically analyze what was happening in the shown lessons and were modeled after prior investigations (e.g., Sherin and van Es 2009, Santagata and Angelici 2010, Kaiser et al. 2015). Pilot versions of our lesson-analysis survey tested both general prompts such as, "What three things stood out to you in this lesson?" and prompts that were specific to a scene in a video, such as, "The instructor asks students to come down to the front of the room and project their work to the whole class. Why do you think she uses this strategy?"

We tested Pilot Version 1 of the lesson-analysis survey by collecting responses from instructors with varying levels of teaching experience and training. Our pilot testing revealed that the more specific prompts elicited shorter answers that revealed less about the knowledge of the respondent. We also saw redundant answers in response to three more general prompts. Lastly, we noticed that not all instructors provided reasoning for their thinking. Therefore, we decided to use two general prompts, and to add explicit instructions to each question to encourage respondents to state their reasoning. Pilot Version 2 of the lesson-analysis survey used different prompts for each video. Though the prompts varied by video, they shared similarities. Prompts included: "Explain what strategies you saw this instructor use that you would keep. Explain the reasons behind your choices."; "What missed opportunities, if any, did you see in this lesson? Explain your reasoning."; and "What did the instructor do that was particularly effective? Explain your reasoning." We further emphasized our requests for reasoning using underlined and italicized text.

Next we used expert feedback to refine the videos and prompts. This approach is commonly used to improve the content validity of the inferences that can be drawn from an assessment. Content validity establishes that the questions appropriately represent the intended knowledge domain (Campbell and Nehm 2013). We sought feedback from researchers who had previously developed and used video-based surveys to assess teacher noticing among K-12 instructors. We asked these experts to draw on their research expertise to provide feedback. We also sought feedback from individuals who work directly with college instructors to provide teaching professional development. We asked these experts to draw on their extensive experience working with faculty to think about how they expected faculty to respond to the

videos and prompts. We gathered feedback from six experts across these two domains using the same questions. We asked them to complete the survey as a respondent would so that they could comment on the whole experience. They also answered four questions for each video: (1) Do you expect this video and prompts to elicit instructor thinking, especially interpretations and evaluations they would make if they were the instructor of the lesson? Why or why not?; (2) Are there prompts, or type of prompts, you would recommend in light of our goal of differentiating experienced from new instructors?; (3) Are there changes to the video that you would suggest?; (4) Do you have any other feedback? We provided experts with incentive for participation and followed-up by email as necessary to fully understand their feedback.

We refined the video-based questions based on this expert feedback to create Pilot Version 3 of the lesson-analysis survey. Expert feedback was overwhelming confirmatory, stating that they expected the videos and writing prompts to elicit teacher knowledge. We made a few changes based on what we learned, including replacing one video. One expert suggested we might elicit more diverse knowledge if some of the videos showed active-learning instruction with more room for improvement, so we added a clip of a more inexperienced instructor. We also eliminated one prompt ("Imagine this instructor has asked you to observe her class and provide feedback. Describe the following and explain your reasoning.") based on expert feedback that placing respondents in the role of providing evaluative feedback might cause them to be more careful in what they write and therefore omit important thoughts. Based on a suggestion from two people, we added a question that asked respondents to synthesize across all three videos. Finally, we recognized that the order of videos and prompts could affect the answers we elicited. For example, a respondent may notice something they would not otherwise notice in Video 3 because they noticed it in Video 1. The most appropriate way to handle this limitation for our study was to keep the order of videos the same for each respondent to avoid introducing variation in answers that results from the order of items.

We collected data from a sample of respondents using Pilot Version 3. The prompts were similar across videos. For the first two videos, the prompts asked respondents to identify what was effective and what could be improved and why, but were stated in different ways. We used only a prompt about effectiveness for the third video because pilot work and expert feedback demonstrated that people--even experts--struggled to identify room for improvement in this video, rendering the question useless in discriminating among instructors with different levels of

expertise. The sixth prompt asked for synthesis across videos: "Consider the three examples of teaching that you have just seen. What teaching strategies used by one or more of these instructors do you consider most important for student learning in large college biology courses?"

The data collected allowed us to refine the prompts a final time and conduct preliminary data analysis. Respondents included nine experts and seven novices. Initial qualitative analysis revealed that the question that asked for synthesis across videos elicited vague responses compared to other questions, so we cut this question. We also noticed that answers written in complete sentences were easier to analyze because we had to make fewer inferences about what the respondents meant. Therefore, we added explicit instructions to use complete sentences to each prompt. We also simplified the instruction provided prior to the videos. The final change we made to prompts was to use the exact same language in prompts across all three videos. In the final version of the lesson-analysis survey, instructors responded to two written prompts after the first two videos: (1) What was effective and why did you think it was effective?, (2) What needs to be improved and why? How would you make it different? After the third video, we asked instructors to respond to question (1). Both questions ended with this sentence: Please use complete sentences.

Appendix B References

Campbell CE, Nehm RH (2013) A critical analysis of assessment quality in genomics and bioinformatics education research. CBE-Life Sciences Education 12, 530-541.

Kaiser G, Busse A, Hoth J, König J, Blömeke S (2015) About the complexities of video-based assessments: Theoretical and methodological approaches to overcoming shortcomings of research on teachers' competence. International Journal of Science and Math Education 13: 369-387.

Santagata R, Angelici G (2010). Studying the impact of the lesson analysis framework on preservice teachers' abilities to reflect on videos of classroom teaching. Journal of Teacher Education 61, 339-349.

Sherin MG, van Es EA (2008) Effects of Video Club participation on teachers' professional vision. Journal of Teacher Education 60(1), 20-37.

APPENDIX C

Full Online Survey

Key:

Instructions provided within the survey appear in this font.

Questions that participants were asked to answer appear in this font.

You will be shown three short videos of college classrooms. We are interested in your observations of teaching and learning in these classrooms. For each video, we will ask you to answer one or both of the following:

1. What was effective and why do you think it was effective?

2. What needs to be improved and why? How would you do it differently?

[Participants watch video 1]
Consider the teaching and learning in this classroom.

What was effective and *why* do you think it was effective? Please use complete sentences.

What needs to be improved and *why*? How would you do it differently? Please use complete sentences.

[Participants watch video 2]
Consider the teaching and learning in this classroom.

What was effective and *why* do you think it was effective? Please use complete sentences

What needs to be improved and *why*? How would you do it differently? Please use complete sentences.

[Participants watch video 3]
Consider the teaching and learning in this classroom.

What was effective and *why* do you think it was effective? Please use complete sentences.

This section asks you about your own classroom teaching.

Consider the biology topics addressed in each video. Please indicate how often you have taught each topic in a college course.

Topic	I have never taught this topic	I have taught this topic but not in the past three years.	I have taught this topic at least once in the past year.
Human ancestry and phylogeny			
Genetic drift			
Golgi structure and function			

How do you know when learning is occurring in your classroom?

How do you maximize student learning in your classroom?

This section asks questions about your own teaching practices. It should take less than 5 minutes to complete. The next page is the last page of questions.

NOTE TO READERS: This section is from the Teaching Practices Inventory (Wieman and Gilbert 2014)

Please consider the largest undergraduate biology course you have taught and answer the following questions with that class in mind.

Class Features and Activities

What course information is provided to students via hard copy or course webpage? Please select ALL that apply.

- ❑ List of topics to be covered
- ❑ List of topic-specific competencies (skills, expertise) students should be able to achieve (what students should be able to do)

- ❏ List of competencies that are not topic related (e.g., critical thinking, problem solving, etc.)
- ❏ Course goal(s) that concern changing students' attitudes and beliefs (e.g., their interest, motivation, perception of relevance, beliefs about their competencies, beliefs about learning)
- ❏ None of the above

To the best of your estimation, please provide the following information about your course. (Please input a numerical character)

- ❏ Number of students enrolled in a single section of this course
- ❏ Number of times per week class meets
- ❏ Average number of times per class that you pause to ask questions
- ❏ Average number of times per class that students work in small groups to answer a question or solve a problem
- ❏ Average number of times per class that you present students with a scenario, demonstration, or simulation
- ❏ Average number of times per class that you ask students to make a prediction about what will occur in a scenario/demonstration/simulation, then show the result and compare students' predictions to the observed result
- ❏ Average number of times per term (i.e., semester or quarter) that you talk to the class about why material is useful and/or interesting to students' lives and careers

Comments on any of the above.

Please select ALL that occur in this course:

- ❏ Students asked to read/view material on upcoming class session
- ❏ Students read/view material on upcoming class session AND complete assignments or quizzes on it shortly before class or at the beginning of class
- ❏ Reflective activity at end of class periods, e.g. "one minute paper" or similar (students briefly answering questions, reflecting on lecture and/or their learning, etc.)
- ❏ Formal student presentations (oral or poster)
- ❏ None of the above

What fraction of a typical class period do you spend lecturing (e.g., presenting content, deriving mathematical results, presenting a problem solution)?

- ❏ 0-20%
- ❏ 21-40%
- ❏ 41-60%
- ❏ 61-80%

- ❑ 81-100%

Considering the time spent on major topics, approximately what fraction was spent on the process by which the theory/model/concept was developed?

- ❑ 0-10%
- ❑ 11-25%
- ❑ More than 25%

Which methods do you use during class to collect responses in REAL TIME from all students? (Check all that occur in your course):

- ❑ electronic ("clickers") with student identifier
- ❑ electronic anonymous
- ❑ colored cards
- ❑ raising hands
- ❑ written student responses that are collected and reviewed in real time
- ❑ None of the above

How many questions per class are followed by student-student discussion?

How many times per class is a student response system used as a quiz device? (counts for points and no student discussion)

NOTE TO READERS: End of Teaching Practices Inventory Questions

For how long have you used the teaching strategies you reported on this page?
_____ Number of terms (i.e., semesters or quarters)

You did it, this is the last page!

Have you engaged in any of the following professional activities? Please select ALL that apply.

- ❑ Participated in 40+ hours of teaching professional development, such as teaching workshops/seminars, graduate-level courses in teaching and learning, etc.
- ❑ Led teaching professional development, such as teaching workshops/seminars, graduate-level courses in teaching and learning, etc.
- ❑ Published peer-reviewed education research
- ❑ Presented education research at a professional conference

- ❏ Had formal training in education research (not teaching) as a graduate student, postdoc, or faculty member
- ❏ Conducted or been trained to conduct classroom observations using a published protocol, such as RTOP, COPUS, EQUIP, or PORTAAL
- ❏ None of the above

To the best of your knowledge, how many semesters have you taught college courses (excluding any semesters you taught as a graduate student)?
_____ Number of terms (i.e., semesters or quarters)

Please upload your CV here. This is one way we can document the scientific approach you have taken to your teaching.

What is your gender?

- ❏ Female
- ❏ Male
- ❏ Other _____
- ❏ No Response

With which race/race(s) do you most closely identify with? Please select ALL that apply.

- ❏ African American/Black
- ❏ American Indian or Alaskan Native
- ❏ Asian
- ❏ Latino
- ❏ Native Hawaiian or other native Pacific Islander
- ❏ White
- ❏ Other _____
- ❏ No Response

With which ethnicity do you most closely identify with?

- ❏ Hispanic
- ❏ Non-Hispanic
- ❏ No Response

Supplemental Material

CBE—Life Sciences Education

Auerbach *et al.*

APPENDIX A

Interview protocol for screening interviews

I will start by telling you a little bit about our study. Although active learning can lead to greater student learning than lecture alone, these strategies are not always effective. We want to dig a little bit more into this, so we are talking to instructors that use active learning to see how long they have been using these strategies, what these strategies look like in their classroom, and how they know when these strategies are working. We believe there is a difference in the knowledge used to implement active learning that results in the different outcomes. We hope to elicit this difference using a video-survey and apply our findings to designing professional development that is aimed at providing instructors with the knowledge and skills they need to become even more effective.

We are using these phone conversations to screen possible participants before referring them to the video survey. However, we have found that there is rich data given related to the screening process and would like your permission to record and use the responses for qualitative analysis. We hope to add another dimension to the knowledge instructors use by looking at the way they describe their own classrooms and instructional decision-making. We will only use information gathered during this phone call should you be deemed eligible and willing to participate in the video-survey. If you are deemed ineligible or choose not to participate in the video-survey, your records will be permanently deleted.

1. Do you teach any large enrollment biology courses?
2. How long have you been using active learning in those courses?
3. What types of active learning do you use, what does AL look like in your classroom?
4. How do you decide what to use?
5. How do you know when something is or isn't working?
6. Are there other ways you use feedback from students to inform your teaching?

APPENDIX B

Developing and Refining Lesson-Analysis Survey

The lesson-analysis survey created for this study included three videos of real active-learning lessons in a college biology course. Each video was accompanied by writing prompts to elicit teacher noticing. Prior lesson-analysis instruments guided our instrument development. We revised and refined the videos and writing prompts using both pilot data and expert feedback. We describe this process in detail below.

Video production

Videos for this instrument needed to be short to minimize participant burden, but also needed to show a section of a lesson that could stand alone. We also required high-quality audio and video and footage of both the instructor and students. We filmed numerous classrooms to collect sufficient and appropriate footage for our research goals. We used two or three cameras each time we filmed, allowing us to simultaneously capture what the instructor and students were doing. We also recorded audio in the classrooms using two to four microphones, including a lapel microphone to clearly capture the voice of the instructor, two shotgun boom microphones pointed at the students to capture student voices, and the microphone on one or more cameras to capture the collective sound of the classroom. We told students in the filmed courses that the videos would primarily be viewed by college instructors, and that designated parts of the room would not be filmed if they did not wish to appear in video. We produced over eight hours of classroom footage from which we created short (i.e., 3-5 minute) clips for the the survey.

We selected and edited video and audio footage to create clips that showed a variety of teaching decisions and practices. We began this process by identifying sections of footage that would be meaningful on their own and that included multiple practices, such as lecturing, small-group discussion, and whole-class discussion. We then spliced instructor footage with footage of students so that viewers had the opportunity to notice both. We added subtitles as necessary so the viewer always had access to what the instructor and featured students were saying. We also began each video with information about the class to provide context for viewers, including the institution type, course title, class size, and current topic. We used Adobe Premier Pro CC v9.2 for all video and audio editing.

Testing and refining lesson-analysis survey

We paired video clips of active-learning classrooms with prompts to elicit teacher knowledge. Prompts asked respondents to critically analyze what was happening in the shown lessons and were modeled after prior investigations (e.g., Sherin and van Es 2009, Santagata and Angelici 2010, Kaiser et al. 2015). Pilot versions of our lesson-analysis survey tested both general prompts such as, "What three things stood out to you in this lesson?" and prompts that were specific to a scene in a video, such as, "The instructor asks students to come down to the front of the room and project their work to the whole class. Why do you think she uses this strategy?"

We tested Pilot Version 1 of the lesson-analysis survey by collecting responses from instructors with varying levels of teaching experience and training. Our pilot testing revealed that the more specific prompts elicited shorter answers that revealed less about the knowledge of the respondent. We also saw redundant answers in response to three more general prompts. Lastly, we noticed that not all instructors provided reasoning for their thinking. Therefore, we decided to use two general prompts, and to add explicit instructions to each question to encourage respondents to state their reasoning. Pilot Version 2 of the lesson-analysis survey used different prompts for each video. Though the prompts varied by video, they shared similarities. Prompts included: "Explain what strategies you saw this instructor use that you would keep. Explain the reasons behind your choices."; "What missed opportunities, if any, did you see in this lesson? Explain your reasoning."; and "What did the instructor do that was particularly effective? Explain your reasoning." We further emphasized our requests for reasoning using underlined and italicized text.

Next we used expert feedback to refine the videos and prompts. This approach is commonly used to improve the content validity of the inferences that can be drawn from an assessment. Content validity establishes that the questions appropriately represent the intended knowledge domain (Campbell and Nehm 2013). We sought feedback from researchers who had previously developed and used video-based surveys to assess teacher noticing among K-12 instructors. We asked these experts to draw on their research expertise to provide feedback. We also sought feedback from individuals who work directly with college instructors to provide teaching professional development. We asked these experts to draw on their extensive experience working with faculty to think about how they expected faculty to respond to the

videos and prompts. We gathered feedback from six experts across these two domains using the same questions. We asked them to complete the survey as a respondent would so that they could comment on the whole experience. They also answered four questions for each video: (1) Do you expect this video and prompts to elicit instructor thinking, especially interpretations and evaluations they would make if they were the instructor of the lesson? Why or why not?; (2) Are there prompts, or type of prompts, you would recommend in light of our goal of differentiating experienced from new instructors?; (3) Are there changes to the video that you would suggest?; (4) Do you have any other feedback? We provided experts with incentive for participation and followed-up by email as necessary to fully understand their feedback.

We refined the video-based questions based on this expert feedback to create Pilot Version 3 of the lesson-analysis survey. Expert feedback was overwhelming confirmatory, stating that they expected the videos and writing prompts to elicit teacher knowledge. We made a few changes based on what we learned, including replacing one video. One expert suggested we might elicit more diverse knowledge if some of the videos showed active-learning instruction with more room for improvement, so we added a clip of a more inexperienced instructor. We also eliminated one prompt ("Imagine this instructor has asked you to observe her class and provide feedback. Describe the following and explain your reasoning.") based on expert feedback that placing respondents in the role of providing evaluative feedback might cause them to be more careful in what they write and therefore omit important thoughts. Based on a suggestion from two people, we added a question that asked respondents to synthesize across all three videos. Finally, we recognized that the order of videos and prompts could affect the answers we elicited. For example, a respondent may notice something they would not otherwise notice in Video 3 because they noticed it in Video 1. The most appropriate way to handle this limitation for our study was to keep the order of videos the same for each respondent to avoid introducing variation in answers that results from the order of items.

We collected data from a sample of respondents using Pilot Version 3. The prompts were similar across videos. For the first two videos, the prompts asked respondents to identify what was effective and what could be improved and why, but were stated in different ways. We used only a prompt about effectiveness for the third video because pilot work and expert feedback demonstrated that people--even experts--struggled to identify room for improvement in this video, rendering the question useless in discriminating among instructors with different levels of

expertise. The sixth prompt asked for synthesis across videos: "Consider the three examples of teaching that you have just seen. What teaching strategies used by one or more of these instructors do you consider most important for student learning in large college biology courses?"

The data collected allowed us to refine the prompts a final time and conduct preliminary data analysis. Respondents included nine experts and seven novices. Initial qualitative analysis revealed that the question that asked for synthesis across videos elicited vague responses compared to other questions, so we cut this question. We also noticed that answers written in complete sentences were easier to analyze because we had to make fewer inferences about what the respondents meant. Therefore, we added explicit instructions to use complete sentences to each prompt. We also simplified the instruction provided prior to the videos. The final change we made to prompts was to use the exact same language in prompts across all three videos. In the final version of the lesson-analysis survey, instructors responded to two written prompts after the first two videos: (1) What was effective and why did you think it was effective?, (2) What needs to be improved and why? How would you make it different? After the third video, we asked instructors to respond to question (1). Both questions ended with this sentence: Please use complete sentences.

Appendix B References

Campbell CE, Nehm RH (2013) A critical analysis of assessment quality in genomics and bioinformatics education research. CBE-Life Sciences Education 12, 530-541.

Kaiser G, Busse A, Hoth J, König J, Blömeke S (2015) About the complexities of video-based assessments: Theoretical and methodological approaches to overcoming shortcomings of research on teachers' competence. International Journal of Science and Math Education 13: 369-387.

Santagata R, Angelici G (2010). Studying the impact of the lesson analysis framework on preservice teachers' abilities to reflect on videos of classroom teaching. Journal of Teacher Education 61, 339-349.

Sherin MG, van Es EA (2008) Effects of Video Club participation on teachers' professional vision. Journal of Teacher Education 60(1), 20-37.

APPENDIX C

Full Online Survey

Key:

Instructions provided within the survey appear in this font.

Questions that participants were asked to answer appear in this font.

You will be shown three short videos of college classrooms. We are interested in your observations of teaching and learning in these classrooms. For each video, we will ask you to answer one or both of the following:

1. What was effective and why do you think it was effective?

2. What needs to be improved and why? How would you do it differently?

[Participants watch video 1]
Consider the teaching and learning in this classroom.

What was effective and *why* do you think it was effective? Please use complete sentences.

What needs to be improved and *why*? How would you do it differently? Please use complete sentences.

[Participants watch video 2]
Consider the teaching and learning in this classroom.

What was effective and *why* do you think it was effective? Please use complete sentences

What needs to be improved and *why*? How would you do it differently? Please use complete sentences.

[Participants watch video 3]
Consider the teaching and learning in this classroom.

What was effective and *why* do you think it was effective? Please use complete sentences.

This section asks you about your own classroom teaching.

Consider the biology topics addressed in each video. Please indicate how often you have taught each topic in a college course.

Topic	I have never taught this topic	I have taught this topic but not in the past three years.	I have taught this topic at least once in the past year.
Human ancestry and phylogeny			
Genetic drift			
Golgi structure and function			

How do you know when learning is occurring in your classroom?

How do you maximize student learning in your classroom?

This section asks questions about your own teaching practices. It should take less than 5 minutes to complete. The next page is the last page of questions.

NOTE TO READERS: This section is from the Teaching Practices Inventory (Wieman and Gilbert 2014)

Please consider the largest undergraduate biology course you have taught and answer the following questions with that class in mind.

Class Features and Activities

What course information is provided to students via hard copy or course webpage? Please select ALL that apply.

- ❑ List of topics to be covered
- ❑ List of topic-specific competencies (skills, expertise) students should be able to achieve (what students should be able to do)

- ❏ List of competencies that are not topic related (e.g., critical thinking, problem solving, etc.)
- ❏ Course goal(s) that concern changing students' attitudes and beliefs (e.g., their interest, motivation, perception of relevance, beliefs about their competencies, beliefs about learning)
- ❏ None of the above

To the best of your estimation, please provide the following information about your course. (Please input a numerical character)

- ❏ Number of students enrolled in a single section of this course
- ❏ Number of times per week class meets
- ❏ Average number of times per class that you pause to ask questions
- ❏ Average number of times per class that students work in small groups to answer a question or solve a problem
- ❏ Average number of times per class that you present students with a scenario, demonstration, or simulation
- ❏ Average number of times per class that you ask students to make a prediction about what will occur in a scenario/demonstration/simulation, then show the result and compare students' predictions to the observed result
- ❏ Average number of times per term (i.e., semester or quarter) that you talk to the class about why material is useful and/or interesting to students' lives and careers

Comments on any of the above.

Please select ALL that occur in this course:

- ❏ Students asked to read/view material on upcoming class session
- ❏ Students read/view material on upcoming class session AND complete assignments or quizzes on it shortly before class or at the beginning of class
- ❏ Reflective activity at end of class periods, e.g. "one minute paper" or similar (students briefly answering questions, reflecting on lecture and/or their learning, etc.)
- ❏ Formal student presentations (oral or poster)
- ❏ None of the above

What fraction of a typical class period do you spend lecturing (e.g., presenting content, deriving mathematical results, presenting a problem solution)?

- ❏ 0-20%
- ❏ 21-40%
- ❏ 41-60%
- ❏ 61-80%

- ❑ 81-100%

Considering the time spent on major topics, approximately what fraction was spent on the process by which the theory/model/concept was developed?

- ❑ 0-10%
- ❑ 11-25%
- ❑ More than 25%

Which methods do you use during class to collect responses in REAL TIME from all students? (Check all that occur in your course):

- ❑ electronic ("clickers") with student identifier
- ❑ electronic anonymous
- ❑ colored cards
- ❑ raising hands
- ❑ written student responses that are collected and reviewed in real time
- ❑ None of the above

How many questions per class are followed by student-student discussion?

How many times per class is a student response system used as a quiz device? (counts for points and no student discussion)

NOTE TO READERS: End of Teaching Practices Inventory Questions

For how long have you used the teaching strategies you reported on this page?
_____ Number of terms (i.e., semesters or quarters)

You did it, this is the last page!

Have you engaged in any of the following professional activities? Please select ALL that apply.

- ❑ Participated in 40+ hours of teaching professional development, such as teaching workshops/seminars, graduate-level courses in teaching and learning, etc.
- ❑ Led teaching professional development, such as teaching workshops/seminars, graduate-level courses in teaching and learning, etc.
- ❑ Published peer-reviewed education research
- ❑ Presented education research at a professional conference

- ❏ Had formal training in education research (not teaching) as a graduate student, postdoc, or faculty member
- ❏ Conducted or been trained to conduct classroom observations using a published protocol, such as RTOP, COPUS, EQUIP, or PORTAAL
- ❏ None of the above

To the best of your knowledge, how many semesters have you taught college courses (excluding any semesters you taught as a graduate student)?
_____ Number of terms (i.e., semesters or quarters)

Please upload your CV here. This is one way we can document the scientific approach you have taken to your teaching.

What is your gender?

- ❏ Female
- ❏ Male
- ❏ Other _____
- ❏ No Response

With which race/race(s) do you most closely identify with? Please select ALL that apply.

- ❏ African American/Black
- ❏ American Indian or Alaskan Native
- ❏ Asian
- ❏ Latino
- ❏ Native Hawaiian or other native Pacific Islander
- ❏ White
- ❏ Other _____
- ❏ No Response

With which ethnicity do you most closely identify with?

- ❏ Hispanic
- ❏ Non-Hispanic
- ❏ No Response