Evaluating the impact of a classroom simulator training on graduate teaching assistants' instructional practices and undergraduate student learning

Tong Wan,[1,2] Constance M. Doty,[2] Ashley A. Geraets,[3] Christopher A. Nix,[3] Erin K. H. Saitta,[3] and Jacquelyn J. Chini[2,*]

[1]*Department of Physics, Westminster College, 1840 South 1300 East, Salt Lake City, Utah 84105, USA*
[2]*Department of Physics, University of Central Florida, 4111 Libra Drive, Orlando, Florida 32816, USA*
[3]*Department of Chemistry, University of Central Florida, 4111 Libra Drive, Orlando, Florida 32816, USA*

(Received 9 July 2020; accepted 10 May 2021; published 29 June 2021)

In this study, we evaluate the impact of rehearsing teaching skills in a mixed-reality classroom simulator on graduate teaching assistants' (GTAs) instructional practices as well as undergraduate student learning outcomes. The simulator training is intended to provide GTAs opportunities to deliberately practice essential pedagogical skills that support active learning, specifically in the context of the combined tutorial and laboratory sections of an algebra-based introductory physics sequence. Over three semesters, GTAs participated in different numbers of simulator rehearsal sessions: no simulator training, one session, and four sessions. We conducted 109 classroom observations for 23 GTAs, using a modified version of the Laboratory Observation Protocol for Undergraduate STEM (LOPUS); we also documented the frequencies of questioning-related skills (e.g., cold calling) implemented by the GTAs. Undergraduate student learning outcomes were measured by pre- and posttests of the Force Concept Inventory (FCI) and Conceptual Survey of Electricity and Magnetism (CSEM). To classify and characterize GTAs' instructional practices, we conducted a hierarchical cluster analysis and found three instructional styles: the small-group facilitator, the whole-class facilitator, and the waiter. The results suggest that four-session simulator training throughout a semester supported GTAs (i) to shift away from the style of the waiter toward the whole-class facilitator, and (ii) to implement posing questions and cold calling techniques. While new GTAs were found to have more interactive behaviors than experienced GTAs in the semester with no simulator training, we found that four-session simulator training supported both new and experienced GTAs to use more interactive instructional styles and to implement questioning-related skills more frequently. Although the results demonstrate the effectiveness of simulator training, our analysis also indicates areas for improvement. GTAs tended to shift away from the style of the small-group facilitator toward the whole-class facilitator when they participated in four-session training, and the weekly implementations of questioning-related skills decreased over the course of a semester despite an increased total implementation. In addition, student learning outcomes in different semesters (with different numbers of simulator rehearsal sessions) did not show a statistically significant difference. However, GTAs' instructional styles were correlated with student performance on FCI posttest with a small effect size when controlling for FCI pretest scores and lecture instructors; no correlation was found between GTAs' instructional style and student performance on the CSEM posttest. We conclude with a discussion of factors that may have led to the success of the simulator training as well as strategies to further enhance the effectiveness of the simulator training.

DOI: 10.1103/PhysRevPhysEducRes.17.010146

I. INTRODUCTION

Graduate teaching assistants (GTAs) play an integral role in undergraduate education in science, technology, engineering, and mathematics (STEM). In many STEM courses at large research universities, GTAs serve as the sole instructors in small sections of recitation and laboratory [1]. Such environments are ideal for active learning as students are given opportunities to work collaboratively in groups and to engage in hands-on activities. Studies have shown that an active learning approach, compared to "teaching by telling," leads to higher performance and lower failure rates in STEM [2]. Research has also demonstrated a positive correlation between GTAs'

*jchini@ucf.edu

teaching skills and student learning outcomes in student-centered active learning environments [3–9].

Although GTAs' content knowledge and implementation of the intended pedagogy can lead to improved student learning outcomes, the extent to which GTAs implement active learning strategies varies substantially [10–16]. Teaching behaviors varied among GTAs even when GTAs participated in extensive training [16]. The results suggest that it is challenging to develop an effective GTA professional development program to support GTAs to engage students in active learning environments.

GTA training programs typically make use of seminars and didactic workshops [17]. Other programs use strategies such as weekly preparation meetings [18] and full courses [19]. These training formats provide GTAs opportunities to experience research-based instructional materials, to review sample student responses, and to learn specific pedagogical skills. The Colorado Learning Assistant (LA) model has been adopted by numerous institutions to prepare undergraduate students for high quality teaching [20]. A key feature of the LA model is experiential learning, where LAs develop pedagogical skills through facilitating student group work in recitation sections and the lecture setting. However, the LA model as well as many other training programs often do not facilitate deliberate practice (i.e., effortful, targeted, and repeated practice [21]) and feedback necessary for improved teaching skills. Moreover, novice LAs or GTAs, who are still in the process of gaining pedagogical content knowledge [(PCK); i.e., integration of pedagogical knowledge and knowledge of subject matter] [22], may not successfully facilitate classroom discourse, which potentially has a negative impact on student learning. Therefore, iterative practice and targeted feedback before classroom practice are necessary for high quality teaching.

Practice-based teaching frameworks are widely adopted in K-12 teacher training [23] and have been shown effective at supporting teachers' classroom practices [24]. During practice-based training, novice teachers are engaged in deliberate practice of specific teaching skills; they are provided targeted feedback and coaching [25]. As iterative practice and feedback are necessary for learners to improve their state of knowledge, deliberate practice is essential for novice teachers to develop teaching skills. Many of the training programs make use of role play, in which one of the participants takes on the role of "teacher" and the others play "students." The Rutgers physics teacher preparation program, for example, makes use of microteaching [26] to support preservice teachers to develop PCK [27]. Becker *et al.* [28] adopted practiced-based teaching in training biology GTAs. They found that GTAs implemented many of the target skills in classroom practice with high frequency, but the implementations were not stable [28]. A potential drawback of role play is that students' states of knowledge, reasoning skills, and behaviors may not be accurately characterized. The lack of authentic teaching experience may reduce the effectiveness of the training [29].

Recently, virtual classroom simulation has been incorporated in teacher preparation. Researchers have investigated the advantages and disadvantages of simulation [30–32]. Simulation provides opportunities to explore and practice in a safe environment before teaching in actual classrooms. Simulated students' states of knowledge can be set such that they closely resemble real students' states. For the aspects we wish to study, simulation affords similar teacher-student interactions as what teachers experience in real classrooms. In addition, simulation facilitates iterative practice and reflection. Simulated students, unlike real students, can be reset to an earlier state, allowing the teacher to retry a lesson segment. Teachers have opportunities to rehearse specific pedagogical skills deliberately, reflect on their rehearsals, and receive targeted feedback. Prior research has demonstrated a positive impact of rehearsal in a mixed-reality simulator on math and science teachers' teaching practices in both simulated environments and real classrooms [29,33–36]. Although simulation has limitations (e.g., limited reality and complexity as well as technical problems [31]), it can be implemented in GTA training to facilitate deliberate practice and feedback. A pilot study has explored the utility of a mixed-reality classroom simulator in an LA program [37]. Chini, Straub, and Thomas found that the classroom simulator created a safe and effective environment for LAs to practice a variety of teaching skills [37].

Informed by practiced-based teaching and the previous success of the mixed-reality classroom simulator, this study explores an effective GTA training model of rehearsing evidence-based teaching practices (i.e., teaching practices that are supported by research) with simulated students. The simulator training supplements weekly GTA preparation meetings, in which the content of curriculum is the primary focus. The simulator training is intended to provide a safe environment for deliberate practice before GTAs feel confident in leading active learning environments. During the simulator training, GTAs are given opportunities to deliberately practice target pedagogical skills, reflect on their use of those skills, and receive feedback from facilitators.

In this paper, we evaluate the simulator training GTAs participated in, drawing on the framework for evaluating GTA PD programs developed by Reeves *et al.* [38]. Reeves *et al.* conceptualized the relationships among contextual variables (e.g., GTA training design), moderating variables (e.g., program adherence), and outcome variables (e.g., undergraduate student performance). They suggested three categories for the outcome variables: GTA cognition, GTA teaching practice, and undergraduate student variables. GTA cognition refers to GTAs' knowledge, skills, attitudes toward or beliefs about teaching. GTA teaching practice refers to GTAs' behaviors regarding planning, instruction, and assessment. Undergraduate student outcomes involve

FIG. 1. Simulator training can impact GTAs' instructional practices, which can in turn impact undergraduate student learning.

gains in knowledge and skills, retention, and interest. The three categories of outcome variables are considered to be related linearly. That is, GTA PD directly impacts GTA cognition, which subsequently influences GTA teaching practice, which then affects undergraduate student outcomes.

Contextual variables include GTA training design, institutional variables, and GTA characteristics [38]. Among these, GTA training design variables (e.g., content, structure, and activities) are deemed to have the most impact on the outcomes. Both institutional variables (e.g., institution type, student population) and GTA characteristic variables (e.g., attitudes toward teaching, prior teaching experience) can affect GTA training design variables. Research shows that GTAs who participated in the same PD program can vary substantially in their cognition and teaching practice [10–16,39]. Moderating variables can influence the relationship between contextual variables and outcome variables. Examples include implementation variables, such as program adherence, and exposure to activities (i.e., dosage).

In line with the framework from Reeves *et al.*, we measure GTA teaching practice and undergraduate student learning to evaluate our GTA PD program. Specifically, we explore the impact of simulator training on GTAs' classroom practices, as well as the subsequent impact on undergraduate student learning (see Fig. 1). In addition, we explored the relationship between GTA teaching experience and the impact on GTAs' instructional practices. We conducted 109 classroom observations for 23 GTAs over three semesters and collected 921 undergraduate student pretest and posttest responses on concept inventories in an introductory physics sequence. In this paper, we address five research questions:

(i) What is the impact of the simulator training on GTAs' instructional styles in actual classrooms?
(ii) What is the impact of the simulator training on GTAs' implementation of questioning-related pedagogical skills in actual classrooms?[1]
(iii) Does GTA prior teaching experience affect the impact of the simulator training on GTAs' classroom practices?
(iv) Through influencing GTAs' classroom practices, what is the subsequent impact of the simulator training on undergraduate student learning outcomes?
(v) How many sessions of simulator training is more effective at supporting GTAs to make use of interactive instructional styles and to implement questioning-related pedagogical skills: one session or four sessions?

[1]We decided to explore this aspect because GTAs were tasked to practice questioning-related pedagogical skills during the simulator training sessions.

II. INTEGRATING SIMULATION IN GTA PROFESSIONAL DEVELOPMENT

A. Instructional context

The study was conducted at a very large, research-intensive metropolitan university in the southeastern United States. Approximately 55% of the students enrolled (including both undergraduate and graduate students) are female and 45% are male. The five most prevalent racial and ethnic groups are White (47%), Hispanic/Latino (27%), Black (11%), Asian (6%), and international (4%).

The target courses involve the two semesters of an introductory physics sequence intended for students who major in life sciences. The first course (Physics I) mainly covers topics in mechanics, and the second (Physics II) covers electricity and magnetism, circuits, and optics. Each course has two components: lecture and a combined tutorial and laboratory session ("mini-studio" [40]). The lecture component is taught by a faculty member, and each mini-studio section is led by a GTA. Typically, each course has two (or three) lecture sections, each with class enrollment of 250–300 students. Each lecture section was accompanied by approximately 9 sections of mini-studio, and each mini-studio is enrolled by up to 32 students.

The weekly mini-studio is comprised of a 75-min tutorial based on the University of Maryland Open Source Tutorials [41], a 15-min group quiz, and an 80-min lab based on the Investigative Science Learning Environment (ISLE) curriculum [42]. During this study, the number of GTAs who taught mini-studios in each semester ranged from 13 to 16. Each GTA typically teaches three sections. In the fall semester of each academic year, approximately half of the GTAs are new (i.e., have not taught either mini-studio or lab before). In the first two semesters of the study, all the mini-studio activities occurred in the instructional laboratory room. In the third semester, however, the tutorial activities for Physics I were moved to a traditional lecture classroom due to increased enrollment and limited laboratory rooms; all the activities for Physics II remained in the laboratory room. The length of the activities for Physics I was adjusted accordingly: the tutorial became 50 min and the lab became 95 min.

B. Pre-existing GTA training in the department

Before the study, the physics department required GTAs to participate in a pedagogy seminar and weekly preparation meetings. All first-year GTAs participate in a semester-long weekly pedagogy seminar, led by a faculty member. During the seminar, GTAs discuss education literature,

evaluate research-based changes to courses, and reflect on their own teaching. All GTAs attend weekly 90-min preparation meetings led by an experienced GTA and/or a postdoctoral researcher. During the meetings, GTAs work through the activities in small groups as students would. They are asked to articulate their answers and reasoning to the tutorial questions; they then design experiments, collect data, and conduct analysis. In addition, GTAs discuss common student difficulties with specific concepts, common student practices in inquiry-based lab activities, and strategies to facilitate student learning.

C. Simulator training

1. Simulator setting

During this study, simulator training was implemented to complement the preexisting departmental-level GTA professional development. The classroom simulator, TeachLivE, combines virtual reality with the physical world [29]—GTAs are physically present in the classroom (without immersion into virtual reality) and interact with student avatars in a virtual classroom that is projected on a large screen (as shown in Fig. 2). The virtual classroom resembles an instructional science laboratory. Five student avatars are divided into two groups to model group work in a laboratory. The student avatars have a range of personalities. Their behaviors are enacted by a trained professional using human-in-the-loop technology, creating an authentic and interactive simulation. During the interaction with student avatars, GTAs wear a portable microphone so that their voice can be captured by the simulator; GTAs' movements are tracked by a motion-sensing input device (Kinect^{TM}).

In each simulator session, GTAs rehearsed target pedagogical skills in a small-group (two or three GTAs) setting. The simulator sessions were facilitated by researchers in discipline-based education. In the first simulator session in each semester, each group was given about 5 min to get familiar with the virtual classroom and student avatars prior to the rehearsals. GTAs then took turns leading a 7-min (approximately) discussion with the student avatars. After the group completed their first rounds of discussion, they reflected on their performance and received feedback from the facilitators. Then, each GTA was given another 7 min to lead a discussion. Lastly, the group concluded the activity with another round of reflection and feedback.

Prior to attending the simulator session, GTAs received instructions for the simulator activity via email. The instructions included an overview of the target pedagogical skills and a physics activity. We reviewed relevant literature [28,43–45] and chose pedagogical skills that have a high impact on student learning. We then adapted activities from the curriculum used in mini-studios so that the activities were suitable for practicing specific pedagogical skills. We also reviewed the literature for common student ideas related to the physics topics and assigned these ideas to the student avatars. The physics activities, suggested student avatars' ideas, and pedagogical goals were provided to the simulator team (See Supplemental Material [46] for an abbreviated example activity). In a pilot study, experienced GTAs were recruited to complete the simulator sessions and to provide feedback as to whether the activities provided opportunities to practice specific pedagogical skills, and whether the simulator provided an authentic teaching experience. As the pilot study progressed, the simulator activities were revised and tested with more experienced GTAs (see Ref. [47] for details of the same project in a chemistry course).

2. Target pedagogical skills

The target pedagogical skills included cold calling, error framing, stretch-it questioning, and group facilitation. As described in Sec. III, GTAs rehearsed different sets of skills in different semesters.

Cold calling and error framing: Cold calling is calling on students to answer a question or participate in a discussion, regardless of whether they have volunteered to do so (e.g., by raising their hand) [43]. Research has shown that frequent cold calling increases student comfort in participation and volunteer participation [45], to promote attendance and engagement [48], and to increase

FIG. 2. A virtual classroom with five student avatars (on the left), and a researcher (acting as a GTA) in a physical classroom (on the right) interacting with the student avatars that are projected on a big screen.

self-reported preparation [49]. Knight, Wise, and Sieke found that using cold call at the group level (e.g., cold calling a group rather than an individual student) correlated with increased discussion of student reasoning and questioning within student groups' discussions [50]. It has been our experience that GTAs find it challenging to facilitate learning when students do not want to participate in discussion voluntarily. Cold calling was therefore chosen as a means for GTAs to promote participation and to increase student comfort.

We note that cold calling is a controversial pedagogical practice. For example, Cooper, Downing, and Brownell demonstrated in an exploratory interview analysis that cold calling may lead to increased student anxiety in a large-enrollment introductory biology course [51]. While the GTAs in this study were working in low-enrollment mini-studios, researchers have found a similar trend among students in lower enrollment community college biology courses [52]. Cooper, Downing, and Brownell propose that cold call increases student anxiety through the mechanism "fear of negative evaluation," the fear of being negatively evaluated that can arise while participating or anticipating participation in a social activity [51,53,54]. To mitigate the fear of negative evaluation, GTAs were tasked with pairing error framing with cold calling. Error framing involves framing mistakes or misconceptions as natural or beneficial [55]. Examples of error framing include telling students "errors are a positive part of the training process" and "you can learn from mistakes and develop a better understanding" [55]. Research has shown that error framing can decrease student anxiety about making mistakes [55] as well as increase student motivation [56] and improve connections between students and faculty [57]. For example, Downing *et al.* state that "instructors' responses to student answers have the potential to significantly decrease students' fear of negative evaluation during whole-class discussion" [52]. Additionally, GTAs were instructed to cold call *after* students had time to discuss their responses with their group, a strategy sometimes renamed "warm calling," which may reduce student feelings of anxiousness by taking the burden for a negative evaluation off the individual student, providing additional time to think about a problem, and providing the opportunity to hear other students' responses [52,58]. Moreover, GTAs were instructed that they may call on either individual students or groups of students after group work.

Stretch it: Questioning is a useful strategy to promote active learning. Koenig, Endorf, and Braun, for example, showed that Socratic questioning implemented by physics GTAs in tutorials has the largest effect on students' conceptual understanding compared to other teaching strategies [5]. Since Socratic questioning is rather complex, we focused on a specific type of questioning technique, called "stretch it." Stretch it is asking students "to explain their thinking" (i.e., stretch it: explain logic) or apply knowledge in new ways (i.e., stretch it: follow-up) [43]." Examples of stretch it: explain logic include "Why do you think that?" and "How do you know that?" Stretch it: explain logic can support students' accountability and logic development [28] and has been shown effective at improving student learning [59,60]. Stretch it: follow-up refers to asking related follow-up questions to stretch boundaries of knowledge and check for integration [28]. For example, "Is there another way to solve this problem?" Stretch it: follow-up can improve students' accountability and higher-order thinking (e.g., application, and synthesis) [28,43].

Group facilitation: Working in groups has been the norm in laboratories and tutorials as numerous studies have demonstrated group work can lead to increased student achievement and improved student attitudes toward science [61–63]. Groups that function well promote learning through engaging with different perspectives, building off others' ideas, and critiquing arguments [64]. However, issues can arise when, for example, a group member is dominating the conversation or reluctant to participate for a fear of being negatively evaluated by peers [51,65]. The student avatars in the simulator were set to have disagreements (e.g., a student avatar dominates the conversation when the group members have different ideas) within their groups. GTAs were tasked with group facilitation, making sure every student's idea is being valued and disagreements are being discussed.

III. METHODS

A. Conceptual framework

The design of this study aligns with the framework developed by Reeves *et al.* [38]. As discussed previously, Reeves *et al.* argue that GTA PD directly impacts GTA cognition, which subsequently influences GTA teaching practice, which then affects undergraduate student outcomes. In line with the framework, we measure GTA teaching practice and undergraduate learning outcomes to evaluate our GTA PD program. Through classroom observations, we classify and characterize GTAs' instructional practices, and measure the extent to which GTAs implemented specific pedagogical skills. In order to explore the effectiveness of the simulator training, we varied the GTA training design and implementation variables. Specifically, we varied the content and activity variable (i.e., rehearsing pedagogical skills with simulator vs no simulator training). We also varied the training dosage (i.e., one-session simulator training vs four session). In addition, we interpret our results in light of contextual variables, such as institutional variables and GTA characteristic variables.

B. Study design and participants

In order to measure the impact of the simulator training, we collected data in three semesters: a fall (semester 0) and a spring semester (semester 1) in one academic year, and another fall term (semester 2) in the following academic

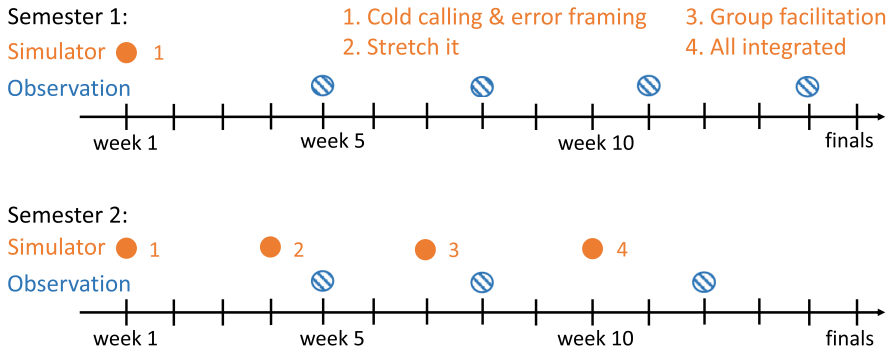

FIG. 3. Timelines of the simulator training and classroom observations in semester 1 and semester 2.

year. In each of the semesters, we conducted classroom observations. We aggregated observation data from both Physics I and Physics II for statistical analysis since the numbers of GTAs in individual courses were small. Therefore, we intended to evaluate the impact on GTAs' classroom practices in the introductory sequence as a whole rather than in individual courses. We also collected undergraduate student responses on concept inventories [Force Concept Inventory in Physics I and Conceptual Survey of Electricity and Magnetism in Physics II] at the beginning and the end of the semester.

We varied the number of simulator training sessions that the GTAs participated in, as shown in Fig. 3. In semester 0, GTAs did not participate in the simulator training. Therefore, the data collected in semester 0 were used as a baseline. Each GTA was observed 3–5 times toward the second half of the semester.[2] In semester 1, GTAs participated in one session of simulator training in week one; they were asked to rehearse cold calling paired with error framing. Each GTA was observed four times throughout the semester. In semester 2, GTAs participated in four sessions of simulator training, with three weeks between sessions. In the first session, GTAs rehearsed cold calling paired with error framing. In the second session, they rehearsed stretch it questioning. In the third session, they rehearsed group facilitation. In the last session, they were asked to integrate all those skills. The sequence of the pedagogical skills was determined based on the increasing level of complexity. Moreover, stretch it can be used following a cold call; group facilitation can be interweaved with error framing and stretch it. This supported deliberate practice, which allowed GTAs to gradually build up their confidence and expertise. Each GTA was observed three times[3] throughout the semester.

[2]The GTAs were observed toward the second half of the semester rather than throughout the semester due to logistical issues, such as getting IRB approval and recruiting participants.

[3]We were planning to observe the GTAs four times, one after each simulator session. However, the observation following the first simulator training session was cancelled due to an unforeseen natural disaster.

The number of GTAs who participated in each semester is shown in Table I. The GTA participation rate ranges from approximately 60% to 80%. Individual GTAs are labeled with alphabetical letters in order to track GTAs who participated in more than one semester. GTAs who had not taught either mini-studio or a calculus-based lab prior to each semester are also shown. The last column presents the number of observations conducted in each semester.

We also report results from undergraduate student performance. The total number of students enrolled over three semesters in Physics I was 1910, and the total number in Physics II was 1061. In both courses, approximately 63% of the students enrolled completed both the pretest and the posttest. Approximately half of the students who completed both tests (31% of the total students enrolled in Physics I and 35% in Physics II) agreed to participate in this study. In order to explore student performance in relation to GTA instructional practices, we only included students whose GTAs also participated in the study. Thus, the number of student participants in Physics I was 425 (22% of total enrollment), and the number of participants in Physics II was 492 (29% of total enrollment). In an effort to evaluate whether the samples are representative, we compared the pretest and posttest scores of the research participants and department-level data (with approximately 63% of the total enrollment); the results suggest that the performance of research participants was similar to all the students who submitted both pretest and posttest (see Appendix A for detail).

C. Classroom observation and interrater reliability

The observation protocol we used is a modified version of the Laboratory Observation Protocol for Undergraduate STEM (LOPUS) [66], as shown in Tables II and III. The protocol documents GTAs' and students' actions in 2-min intervals. More than one action can be coded in the same 2-min interval. The codes do not rely on predefined criteria for teaching quality. Instead, they describe the behaviors expected of GTAs and students, such as the GTA talking to an individual or group of students (1o1-Talk), and student

TABLE I. Number of GTAs participated in classroom observations in each semester.

	Number of participating GTAs	Participation rate	Individual GTAs participated	Number of new GTAs	Number of observations
Semester 0: no simulator	8	8/13	a–h	5	33
Semester 1: One session	10	10/13	a–g, i–k	1	39
Semester 2: Four sessions	13	13/16	k, l–w	7	37

TABLE II. Definitions for codes that describe GTA behaviors.

Type of behavior	GTA code	Abbreviated definition
Typical instructional behaviors	Lec	Lecturing to the class or making announcements
	RtW	Real-time writing on the board, doc cam, etc.
	FUp	Providing follow-up or feedback on activity
	D/V	Showing a demonstration or video
	M	Monitoring class or individual groups
Interactive behaviors	PQ	Posing a worksheet- or lab-related question
	1o1-Talk	Talking to individual student or group of students one-on-one
	1o1-TPQ	Posing a question to individual students or group of students
	VF	Providing feedback to student responses
	VM	Verbal monitoring
	TI	Initiating one-on-one interaction with students
Noninstructive behaviors	Adm	Performing administrative tasks
	W	Waiting and generally unavailable to students

TABLE III. Definitions for codes that describe student behaviors.

Type of behavior	Student code	Abbreviated definition
Typical instructional behaviors	Wks/Lab	Working on worksheets or performing lab activity
	TQ	Taking a quiz
Interactive behaviors	SQ	Asking the GTA a worksheet- or lab-related question with entire class listening
	1o1-SQ	Individual student or a group of students asking the GTA a worksheet- or lab-related question
	WC	Engaging in whole class discussion in which students are talking serially
	SI	Initiating one-on-one interaction with the GTA

posing a question with the entire class listening (SQ). Prior to the study, three observers conducted practice observations with the original LOPUS. We then examined the face validity of the protocol, considering that our instructional context and research goals differ from the study in which LOPUS was developed. We have the expertise to assess the face validity since one of us led the course development, and two of us led the weekly preparation meetings and/or taught mini-studios before. In order to fit the instructional context and research goals of this study, we added an additional code (GTA provides verbal feedback); we also eliminated codes due to either low occurrence (e.g., student presentation) or the constraint of real-time coding (e.g., student waiting) (see details of modifications and the corresponding reasons in Ref. [67]).

In addition to the codes shown in Tables II and III, we also included codes with respect to cold calling (see Table IV), one of the target pedagogical skills in the simulator training. We did not include codes relevant to error framing and group facilitation because these skills are complex and require in-depth qualitative analysis, which is beyond the scope of this quantitative study. Additionally, we expect these skills will be used infrequently, so investigating the use of error framing and group facilitation is likely a better fit for qualitative analysis. Future research will examine *how* GTAs implement error framing and group facilitation. We also did not include stretch it in this paper since we did not collect baseline data for this skill (i.e., not collected in semester 0 when simulator training was not implemented).

TABLE IV. Definitions for codes with respect to target pedagogical skills.

Pedagogical skill	Abbreviated definition
CC	Cold calling with entire class listening; either individuals or groups of students can be called on
1o1-CC	Cold calling in individual groups of students

In semesters 0 and 1, three observers conducted the classroom observations. In semester 2, a fourth observer joined the observation team.[4] Prior to the study, the three observers conducted practice observations in order to modify the definitions of codes and ensure individual observers were applying codes similarly. Before semester 2, the fourth observer conducted practice observations with the other observers and demonstrated they were coding similarly to the team before formal observations took place.

To explore interrater reliability (IRR), the first three classroom observations in semester 0 and the first two observations in semester 2 were conducted by either pairs or triads of observers. Each of these observations involved a unique GTA. We calculated IRR for pairs of observers using Cohen's kappa for the overall IRR of the modified LOPUS (i.e., IRR for all the codes in the modified LOPUS). We also calculated IRR for individual codes (e.g., IRR for code Lec) to explore agreement on the level of individual behaviors, including the codes related to cold calling. For the IRR of individual codes, we used Gwet's AC1 because it is less sensitive to extreme (i.e., either very large or very small) trait prevalence compared to Cohen's kappa [68]. Both methods involved calculating percent agreement, but they differed in the way that the chance agreement probability is determined. For the overall IRR, the unit for calculating percent agreement is every code in every 2-min interval; for the IRR of an individual code, the unit for calculating percent agreement is every 2-min interval for that specific code. After calculating the IRR for each pair of observers in each observation, we calculated the average IRR over all pairs of observers in all the observations.[5] For all four observers, we achieved an average prediscussion Cohen's kappa of 0.74 ± 0.11, and an average prediscussion Gwet's AC1 of 0.86 ± 0.18. These results suggest that all four observers, on average, were in good agreement (0.61 to 0.80) for the modified LOPUS, and very good agreement (0.81 to 1.00) for individual codes according to Altman's criteria [68].

[4]An additional observer was recruited due to the increased number of total observations that were conducted in physics, chemistry, and math contexts.

[5]The average overall IRR was calculated over all pairs of observers in all the observations; the average IRR for individual codes was calculated over all the codes for all pairs of observers in all the observations.

IV. IMPACT ON GTA CLASSROOM PRACTICES

A. GTAs' instructional style categorized by cluster analysis

1. Cluster analysis

Inspired by Velasco *et al.* [66], we conducted a cluster analysis in R [69] using the modified LOPUS codes to classify and characterize GTAs' instructional styles. Unlike comparing the frequency for each individual LOPUS code, cluster analysis examined the frequencies of all the codes holistically. Therefore, cluster analysis allowed us to evaluate GTA classroom practices holistically. The study from Velasco *et al.* was conducted in a traditional chemistry laboratory. Since the physics mini-studios make use of research-based curriculum and active learning strategies, we expect that GTA and student behaviors in physics mini-studios would be different from a traditional laboratory, and therefore would result in different clusters. In addition, not only did we intend to classify and characterize GTAs' instructional styles, we also aimed to identify trends that GTAs shift between different instructional styles as a result of simulator training. Therefore, cluster analysis was considered suitable for our research purposes.

We used individual class periods observed instead of individual GTAs as the units for analysis since each GTA was observed multiple times. Multiple observations were necessary as suggested by a national study on STEM faculty's teaching practice that faculty may use multiple sets of practices across the observations [70]. We calculated the fraction that each code occurred during the entire class period. For example, if PQ was coded in four 2-min intervals during an observation that included 80 2-min intervals, the fraction would be 0.05 (4/80). We then scaled the fractions for each code across all the observations such that the average is 0 and the standard deviation is 1. This allowed each code to have an equal weight in clustering. Since we were interested in identifying codes in which the clusters differ, the scaling allowed the follow-up statistical tests to tease out codes that did not occur frequently but differed significantly between the clusters.

We used agglomerative hierarchical clustering, specifically Ward's method [71]. Agglomerative hierarchical clustering follows a bottom-up, iterative process. Initially, each class period was in its own cluster. The distances between the clusters were calculated in a Euclidian space defined by all the codes of the modified LOPUS. The two clusters that were most similar (i.e., have the smallest distance) were combined. The process was repeated until all the class periods were in

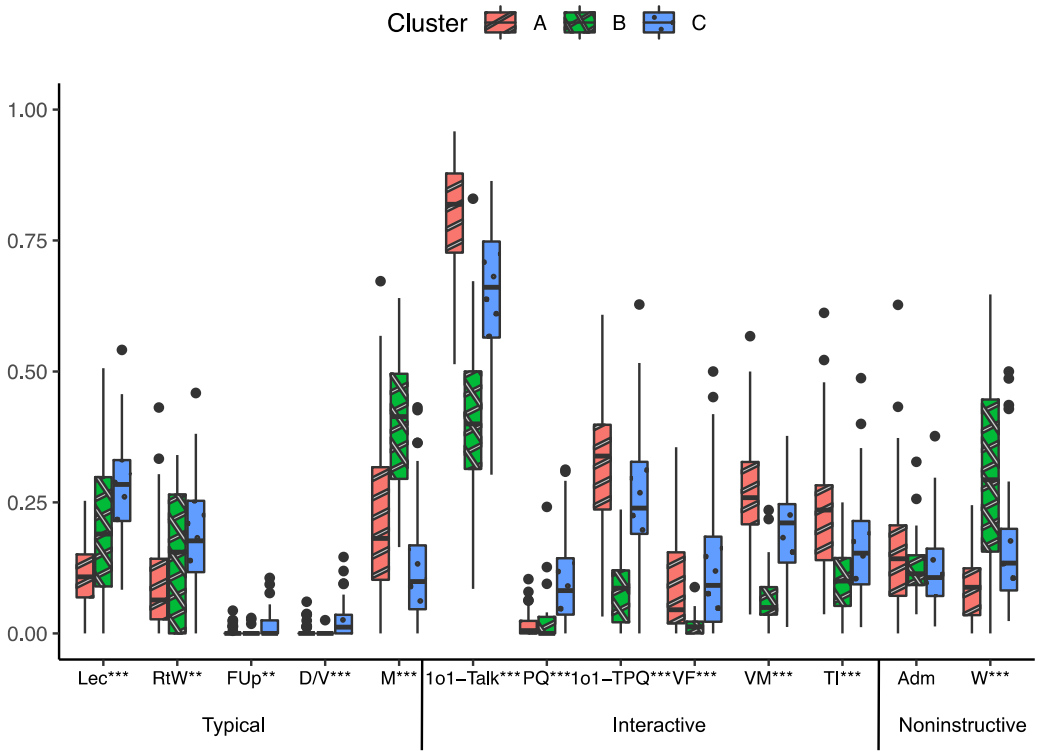

FIG. 4. Fraction of 2-min intervals per class period for each GTA code in each cluster. Out of 13 GTA codes, 12 codes are statistically significant between clusters. **medium effect size, ***large effect size.

one cluster. We then used the "elbow" method [72] and gap statistics [73] with the *factoextra* package [74] to explore the optimal number of clusters.

After the number of clusters was determined, we identified codes that were significantly different between the clusters. Considering our small sample size, we used the Kruskal-Wallis rank sum test [75], a nonparametric test that ranks data before a comparison of the mean ranks. To determine the effect size, we used the eta squared for the Kruskal-Wallis test [76]. In this paper, we report codes that have large effect size ($\eta^2 > 0.11$), and medium effect size ($\eta^2 > 0.06$) [77]. In order to identify which clusters were significantly different from the others, we conducted a *post hoc* analysis using Dunn's multiple comparison test [78] with the *dunn.test* package [79]. To reduce type I error rate due to multiple comparisons, P values were corrected with the Holm-Bonferroni method [80].

2. Cluster analysis results

The hierarchical cluster analysis suggests that the classroom observation data display a cluster structure (See Appendix B, Fig. 12). Both the elbow method and the gap statistics suggest three as the optimal number of clusters. We found two student codes and ten GTA codes that are statistically significantly different between the clusters with large effect sizes, and one student code and two GTA codes with medium effect sizes (see Figs. 4 and 5). The complete results from Kruskal-Wallis tests and Dunn's tests can be found in Appendix B, Table IX. We note that only one GTA code (Administration) did not show a difference among clusters, suggesting that the behaviors of GTAs in different clusters differed in a variety of ways.

3. Characteristics of GTA instructional styles

We used GTA codes with large effect sizes to characterize GTAs' instructional styles, as shown in Table V. For each instructional style, we present students' behaviors (with medium and large effect sizes[6]) to demonstrate the relationships between student behaviors and GTA instructional styles [67]. The three instructional styles we found are the group-work facilitators, the waiters, and the whole-class facilitators. Below we describe these three instructional styles.

The group-work facilitators: GTAs facilitate students working in groups. Among all three clusters, GTAs in cluster A had the highest fractions of talking to individual students or groups of students one-on-one [1o1-Talk, $\chi^2(2) = 55.5$, $p_{\text{K-W}} < 0.001$] and initiating conversations with students [TI, $\chi^2(2) = 20.2$, $p_{\text{K-W}} < 0.001$]; they

[6]We decided to use codes with both medium and large effect sizes instead of large effect sizes only since there are fewer student codes than GTA codes, and the student code with a medium effect size, 1o1-SQ, provides us meaningful insights into how students' behavior can be different across clusters.

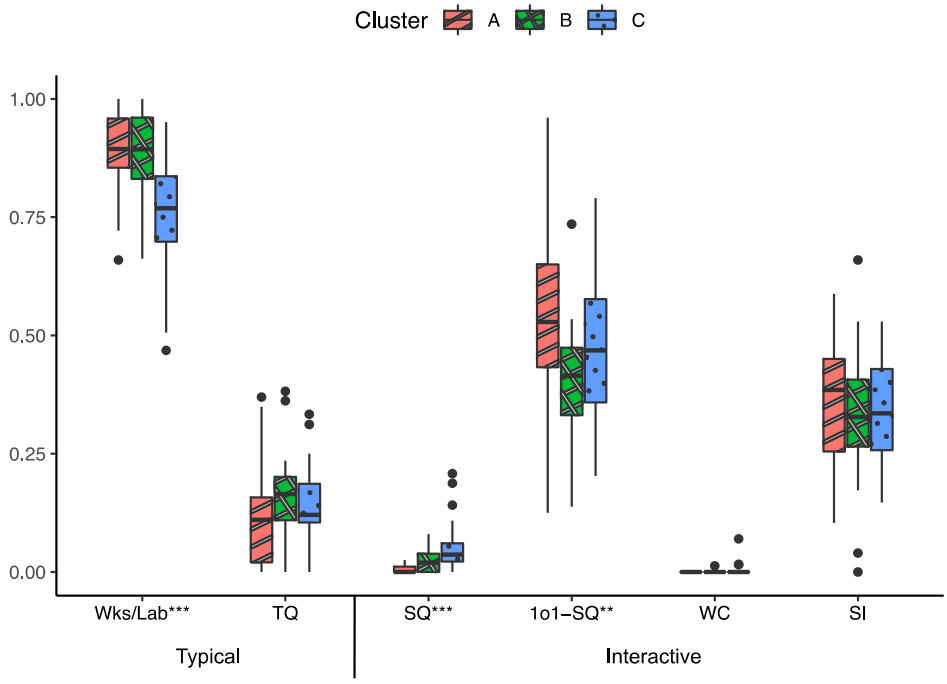

FIG. 5. Fraction of 2-min intervals per class period for each student code in each cluster. Out of 6 student codes, 3 codes are statistically significant between clusters. **medium effect size, ***large effect size.

also posed the most questions in one-on-one interactions [1o1-TPQ, $\chi^2(2) = 37.7$, $p_{\text{K-W}} < 0.001$]. In addition, they spent the least time lecturing [Lec, $\chi^2(2) = 44.3$, $p_{\text{K-W}} < 0.001$] and waiting [W, $\chi^2(2) = 29.9$, $p_{\text{K-W}} < 0.001$]. Correspondingly, students in cluster A asked more questions in one-on-one interactions [1o1-SQ, $\chi^2(2) = 9.5$, $p_{\text{K-W}} = 0.009$, $p_{\text{H-B}} = 0.003$] compared to cluster B; they asked the least questions with the whole class listening [SQ, $\chi^2(2) = 47.8$, $p_{\text{K-W}} < 0.001$] among all clusters.

The waiters: GTAs tend to wait until students call on them and interact with students less frequently. Among all three clusters, GTAs in cluster B spent the most time monitoring [M, $\chi^2(2) = 35.1$, $p_{\text{K-W}} < 0.001$] and waiting (W); they had the lowest fractions of talking to individual students or groups of students one-on-one (1o1-Talk), initiating conversations with students (TI), and posing questions in one-on-one interactions (1o1-TPQ). Moreover, they provided less verbal feedback [VF, $\chi^2(2) = 18.6$, $p_{\text{K-W}} < 0.001$] compared to both cluster A ($p_{\text{H-B}} < 0.001$) and cluster C ($p_{\text{H-B}} < 0.001$). As mentioned previously, students in cluster B asked fewer questions in one-on-one interactions (1o1-SQ) compared to cluster A.

The whole-class facilitators: GTAs engage students in the whole-class setting. GTAs in cluster C had the highest fraction of lecturing (Lec) among all three clusters. They asked more questions with the whole class listening [PQ, $\chi^2(2) = 31.8$, $p_{\text{K-W}} < 0.001$] and spent more time showing a demonstration or video [D/V, $\chi^2(2) = 21.2$, $p_{\text{K-W}} < 0.001$] compared to both cluster A and cluster B. Students in cluster C asked the most questions with the whole class listening (SQ) among all clusters, and spent less time on worksheets or lab activities [Wks/Lab, $\chi^2(2) = 34.9$, $p_{\text{K-W}} < 0.001$] compared to both cluster A ($p_{\text{H-B}} < 0.001$) and cluster B ($p_{\text{H-B}} < 0.001$).

4. Similarities between clusters

We have described the instructional styles based on unique characteristics of each cluster. However, it is also important to point out the similarities between two of the instructional styles: the group-work facilitators and the whole-class facilitators. The group-work and the whole-class facilitators had more interactions with students than the waiters did. As shown in Fig. 4, both the group-work and the whole-class facilitators had drastically higher fractions of 1o1-Talk, 1o1-TPQ, and VM compared to the waiter; they also had much lower fractions of M and W compared to the waiters. Therefore, the waiter is the least desired instructional style among all three. In addition, student behaviors are linked with GTA instructional styles; when GTAs made use of more interactive styles (the group-work facilitator and the whole-class facilitator), students were more engaged as they asked more questions.

B. Correlation between GTA instructional style and simulator training

1. Analysis

In order to answer RQ1 and RQ5, we investigated the correlation between GTA use of instructional style and the

TABLE V. Definitions of GTA instructional styles characterized by codes that have medium and large effect sizes.

Cluster	Instructional style[7]	Behaviors that occurred either more frequently or less frequently than the other two clusters (Median fraction)
A (42 sessions)	The group-work facilitators	*GTAs*: spent the most time talking to students one-on-one (0.82) posed the most questions one-on-one (0.34) initiated the most one-on-one interactions (0.24) spent the least time lecturing (0.11) spent the least time waiting (0.09) *Students*: asked more questions one-on-one* (0.53) asked the least questions with whole class listening (0.00)
B (23 sessions)	The waiters	*GTAs*: spent the most time monitoring (0.41) spent the most time waiting (0.29) spent the least time talking to students one-on-one (0.40) initiated the least one-on-one interactions (0.10) asked the least questions in one-on-one interactions (0.09) provided the least verbal monitoring (0.05) provided less verbal feedback (0.01) *Students*: asked fewer questions in one-on-one interactions* (0.42)
C (44 sessions)	The whole-class facilitators	*GTAs*: asked more questions in front of whole class (0.08) spent more time on demonstration or video (0.01) had the highest fraction of lecturing (0.28) *Students*: asked the most questions with whole class listening (0.04) spent less time on worksheets or lab activities (0.77)

*Only one pair of clusters are distinguishable; the third cluster is not distinguishable from the other two.

semester, during which the number of GTA simulator sessions varied. We performed a chi-square test to compare the proportions of the observations in clusters between different semesters.[8] Effect size was determined by Cramer's V [76]. This was followed by pairwise comparisons with Holm-Bonferroni correction.

2. Results

Table VI shows the distributions of class periods observed in each cluster in each semester. In both semesters 0 and 1, the group-work facilitator was used more frequently than the other two styles. In semester 2, the whole-class facilitator was used much more frequently, and the waiter was used less frequently compared to semesters 0 and 1.

A chi-squared test with all the data in Table VI suggests that GTAs' use of instructional styles depends on the number of simulator sessions the GTA participated in with a medium effect size [$\chi^2(4) = 14.3$, $p = 0.006$, Cramer's $V = 0.256$]. Pairwise comparisons between semesters with Holm-Bonferroni correction show that the difference in proportions came from two pairs: semester 0 and semester 2 [$\chi^2(2) = 9.9$, $p = 0.007$], as well as semester 1 and semester 2 [$\chi^2(2) = 10.5$, $p = 0.005$]. Considering there were three instructional styles and we were interested in how the distributions were different among the semesters, we conducted pairwise comparisons between clusters. The results suggest that the difference was statistically significant between cluster A and cluster C [$\chi^2(2) = 10.0$, $p = 0.007$] as well as cluster B and cluster C [$\chi^2(2) = 8.6$, $p = 0.014$]. This may suggest that the four-session simulator training shifts GTAs from both the group-work facilitators and the waiters toward the whole-class facilitators. However, it may also be due to the difference in the physical learning space [81] of the tutorial activities in

[7]We used the same names of instructional styles as the physics instructional context in Ref. [67], but the characteristics of the styles differ slightly. First, Ref. [67] and this study used almost identical codes to describe the clusters except that RtW was used in Ref. [67] only and M was used in this study only. Second, for three of the codes used in both studies (i.e., TI, VM, and W), all three clusters are distinguishable from each other in this study while only one or two pairs of clusters are distinguishable in Ref. [67]

[8]We chose to examine by semester rather than by number of simulator sessions ranging from 1 to 5 because most GTAs participated in either 1 session in semester 1, or 4 sessions in semester 2. Only one GTA participated in both semester 1 and semester 2, as shown in Table I.

TABLE VI. Distributions of class periods observed in each cluster in different semesters with various number of simulator sessions.

	Cluster A: The group-work facilitators	Cluster B: The waiters	Cluster C: The whole-class facilitators
Semester 0 ($n = 33$) No simulator	16 (48%)	8 (24%)	9 (27%)
Semester 1 ($n = 39$) One-session simulator	17 (44%)	11 (28%)	11 (28%)
Semester 2 ($n = 37$) Four-session simulator	9 (24%)	4 (11%)	24 (65%)

Physics I sections. As mentioned before, in semester 2, the Physics I format was changed so that the tutorial portion took place in a traditional classroom with individual desks rather than the lab room with tables. Since the physical learning space was not a factor in the Physics II sections, we examined the proportions of instructional styles in Physics II. The data suggest a shift in Physics II sections as well. In semester 0, 13 out of 17 class periods of Physics II were categorized in cluster A, four were in cluster B, and zero were in cluster C; in semester 2, seven out of 14 Physics II sections were in cluster C, four were in cluster A, and three were in cluster B (see Appendix C for further analysis of differences in instructional styles between Physics I and Physics II). Therefore, we conclude that the four-session simulator training appears to lead GTAs to make use of the whole-class facilitator more often and the group-work facilitator and the waiter less often.

C. Correlation between GTA implementation of questioning-related skills and simulator training

1. Data source and analysis

To answer RQ2 and RQ5, we explored the correlation between GTA implementation of specific pedagogical skills and GTA simulator training. Specifically, we examined three questioning-related skills: cold calling, posing questions with whole class listening, and posing questions in one-on-one interactions. As mentioned previously, cold calling was rehearsed in the simulator, and the goal of implementing cold calling is to promote student participation. Posing questions with whole class listening and in one-on-one interactions with cold calling together provide insight into how often GTAs used questioning techniques to facilitate student learning.

We used frequency (i.e., number of 2-min intervals) instead of fraction in this analysis since we were interested in how many times the skills were implemented rather than how much of the class time was allocated for questioning. Moreover, we did not expect cold calling to occur very frequently during a single class period. (Ideally, we would like GTAs to implement cold calling a few times in every class period.) Therefore, we decided that frequency is more appropriate to describe skill use. We note that a caveat for reporting frequency is that the duration of a class period may be a factor that influences the frequencies of the skills.[9] We consider the extent to which the duration affects the frequencies of cold calling and posing questions with whole class listening to be low since these skills are usually used toward the first half of a mini-studio (i.e., during the tutorial and the beginning of the lab). However, posing questions in one-on-one interactions, which is expected to occur until the class ends, can be substantially affected by the duration if the durations of classes vary significantly. We will discuss how the variance of class durations may have affected our results.

Using the Kruskal-Wallis test and the eta-squared for the Kruskal-Wallis test, we compared the average frequencies of questioning-related skills in different semesters when GTAs participated in different numbers of simulator sessions. We also examined how the average frequencies change over the course of each semester with the Wilcoxon signed rank test. The effect size was determined by r [76] with the *rcompanion* package [82].

2. GTA implementation of questioning-related skills across semesters

We report the frequencies of cold calling, posing questions with whole class listening, and posing questions in one-on-one interactions. Since individual frequencies of cold calling with whole class listening (CC) and in one-on-one interactions (1o1-CC) were low, we report the sum of the frequencies (CC-total). The frequency of each of the skills in each semester is shown in Figure 6. The frequencies of CC-total and PQ appeared to have an increasing trend. We conducted a Kruskal-Wallis' test for each skill. We found that the frequencies of CC-total [$\chi^2(2) = 12.1$, $p = 0.002$] and PQ [$\chi^2(2) = 28.6$, $p < 0.001$] were statistically significantly different across the semesters, while the difference in frequencies of 1o1-TPQ [$\chi^2(2) = 1.2$, $p = 0.562$] was not statistically significant. The difference for CC-total had a medium effect size [$\eta^2(2) = 0.095$], and it had a large effect size for PQ [$\eta^2(2) = 0.251$]. For both

[9]We also did an analysis with the fractions of the skills. The results from the analyses with fractions and with frequencies were fairly consistent. We were able to draw the same conclusion for comparisons of skill use across semesters. For skill use over the course of a semester, both analyses led us to the same finding for semester 2; for semester 1, both PQ and 1o1-TPQ decreased if we compared the frequencies while only PQ decreased if we compared the fractions.

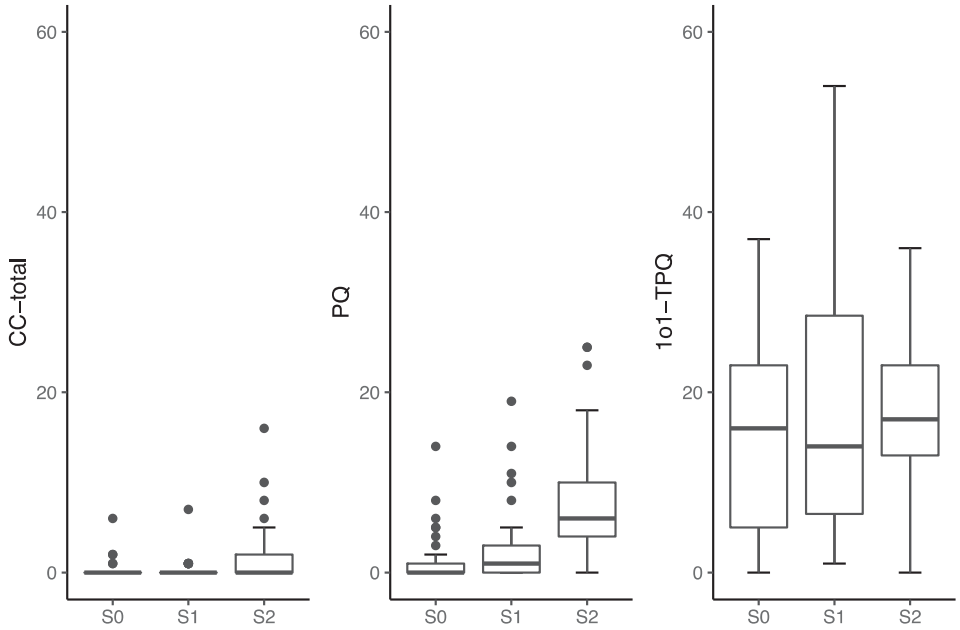

FIG. 6. Frequencies of questioning-related skills across semesters with different numbers of simulator sessions. CC-total is statistically significant with a medium effect size; PQ is statistically significant with a large effect size; 1o1-TPQ is not statistically significant. For both CC-total and PQ, the frequencies in semester 2 are higher compared to both semester 0 and semester 1.

skills, Dunn's test suggests that the differences were statistically significant between semester 0 and semester 2, as well as between semester 1 and semester 2. This seems to suggest that the larger frequencies of skill implementation in semester 2 were due to the four-session simulator training GTAs participated in.

Considering the class duration may also be a factor impacting the frequencies of skill, we examined the average class durations across semesters. The results show that the class durations across three semesters were also significantly different with a large effect size [$\chi^2(2) = 19.7$, $p < 0.001$, $\eta^2(2) = 0.167$]. Dunn's test suggests that the average duration in semester 2 was the largest and the average duration in semester 0 was the smallest. These results suggest that an increase in class duration does not necessarily result in higher frequencies as we did not find a difference in average frequency between semester 1 and semester 0 for any of the three skills. Thus, it is reasonable to believe that the larger frequencies in semester 2 resulted from the simulator training. In addition, 1o1-TPQ, which is most likely to be affected by duration among all three skills, did not show a difference in frequency between semester 2 and the other two semesters. Therefore, we conclude that the four-session simulator training that GTAs participated in during semester 2 supports GTAs' implementation of cold calling and posing question with whole class listening.

3. GTA implementation of questioning-related skills over the course of a semester

We also examined how the frequencies of skills changed over the course of a semester. Figure 7 shows the frequencies of the three skills during each observation in semesters 1 and 2. Data from semester 0 were not included because not every GTA was observed in the same weeks and that not every GTA was observed the same number of times. The frequencies of PQ and 1o1-TPQ in both semesters appear to have a decreasing trend. For ease of interpretation, we only compared the frequencies for the first and last observations in each semester as opposed to including all the observations in the statistical analysis. We used the Wilcoxon signed rank test for matched samples. In semester 1, both PQ ($p = 0.024$, $r = -0.734$) and 1o1-TPQ ($p = 0.014$, $r = -0.734$) in week 14 had decreased average frequencies with large effect sizes compared to week 5. In semester 2, only PQ ($p = 0.016$, $r = -0.707$) in week 12 had a decreased average frequency with a large effect size compared to week 5. The results suggest that although simulator training sessions support GTAs to implement questioning-related skills, GTAs' implementations were not always stable over the course of a semester. Additionally, we also examined the average duration of the mini-studio sessions. We found that, in both semesters, the average durations decreased in the last observations compared to the first observations with large effect sizes. However, the average frequency for 1o1-TPQ in week 12 was comparable to that in week 5 in semester 2. This may also indicate the effectiveness of four-session simulator training.

D. Impact on instructional practices for GTAs with and without prior teaching experience

1. Analysis

To answer RQ3, we compared the proportions of class periods observed in different clusters between new and

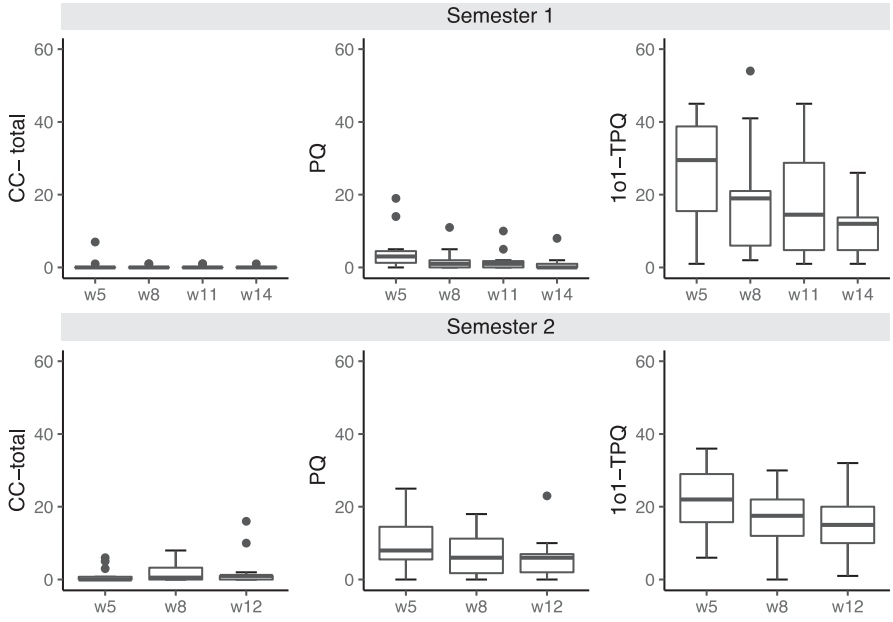

FIG. 7. Frequencies of questioning-related skills over the course of semester 1 and semester 2, respectively. In semester 1, each GTA was observed four times, and the average frequencies of both PQ and 1o1-TPQ were lower in week 14 compared to week 5. In semester 2, each GTA was observed three times, and PQ had a lower average frequency in week 12 compared to week 5.

experienced GTAs. Since the numbers of class periods in different groups were small, we only performed a qualitative comparison. In addition, we compare GTAs with and without prior teaching experience for the frequencies of skills in different semesters to answer RQ3. We performed Wilcoxon rank sum tests with cliff's *d* [83] for effect sizes.

We did not include data from semester 1 for two reasons. First, the numbers of new and experienced (i.e., taught before) GTAs in semester 1 were drastically different (only one out of 10 GTAs was new). Second, our results discussed previously suggest that one-session simulator in semester 1 did not impact the GTAs' instructional practices. Therefore, we compared results from semester 0 and semester 2 to explore whether the impact of four-session simulator training differs between GTAs with and without prior teaching experience.

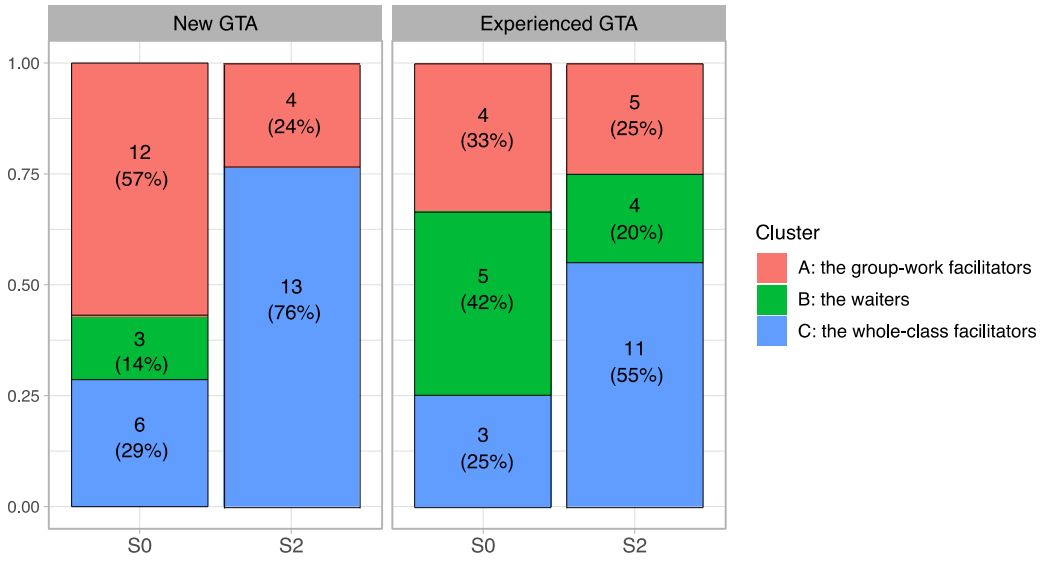

FIG. 8. The percentages of class periods observed in different clusters for GTAs with different teaching experiences. Compared to semester 0, both new and experienced GTAs in semester 2 used the group-work facilitator and the waiter less frequently, but more frequently used the whole-class facilitator. None of the new GTAs in semester 2 used the waiter.

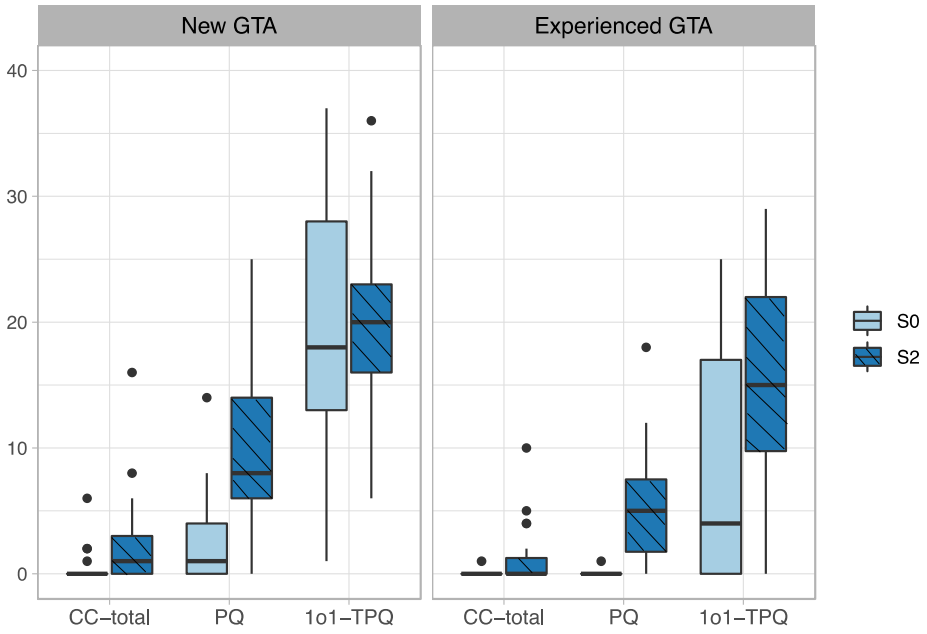

FIG. 9. Frequencies of questioning-related skills implemented by GTAs with and without teaching experience in semester 0 and semester 2. For both new and experienced GTAs, the average frequencies of CC-total and PQ increased in semester 2, with medium and large effect size, respectively.

2. Results

Figure 8 shows the percentages of class periods observed in each cluster that corresponds to different GTA instructional styles for new and experienced GTAs in semester 0 and semester 2. We compare the results descriptively without using statistical tests due to small sample sizes. In semester 0, new GTAs appear to make use of the small-group facilitator more often and the waiter less often compared to experienced GTAs. In semester 2, both new and experienced GTAs shifted away from the small-group facilitator and the waiter toward the whole-class facilitator. Notably, none of the new GTAs in semester 2 used the style of the waiter. The results appear to suggest that both new and experienced GTAs are receptive to PD.

We also investigated changes in instructional styles for the six GTAs who participated in both semester 0 and semester 1. We found that five of the GTAs changed their uses of instructional styles (e.g., a GTA who used the style of the group-work facilitator in semester 0 used both the group-work facilitator and the waiter in semester 1). However, we did not notice a clear trend in the changes, which may be due to a small sample size.

Figure 9 shows the frequencies of questioning-related skills implemented by GTAs with and without teaching experience in semester 0 and semester 2. In semester 0, new GTAs implemented higher frequencies of PQ (Wilcoxon rank sum test, $p = 0.002$) and 1o1-TPQ ($p = 0.011$) compared to experienced GTAs, both with large effect sizes (Cliff's $d = 0.615$ and $d = 0.540$, respectively). In semester 2, both new and experienced GTAs had greater frequencies of CC-total ($p = 0.033$ and $p = 0.049$, respectively) with medium effect sizes ($d = 0.347$ and $d = 0.338$, respectively) compared to semester 0; they also had larger frequencies of PQ ($p < 0.001$ for both) with large effect sizes ($d = 0.734$ and $d = 0.780$) compared to semester 0. In semester 2, only one code, PQ ($p = 0.028$), had higher frequency for new GTAs with a medium effect size ($d = 0.426$) compared to experienced GTAs. The results suggest that although new GTAs implemented questioning-related skills more frequently in the baseline data, the four-session simulator training supported both new and experienced GTAs to implement these skills.

IV. IMPACT ON UNDERGRADUATE STUDENT LEARNING

A. Data source and analysis

We examined how student performance on FCI and CSEM differed across the semesters to explore the correlation between student learning and GTA simulator training, which allowed us to answer RQ4. To control for student incoming preparation (measured by pretest) and lecture instruction, we used ANCOVA (or ANOVA[10]) on student posttest across the semesters. We also explored the correlation between student learning and GTA instructional approach. GTA instructional approach was determined

[10]ANOVA was used for data from Physics II due to the nonlinear relationship between pretest and posttest scores. See Sec. IV, part B for details.

TABLE VII. Number of undergraduate student participants in each condition. The percentage is out of the total number of participants in each semester. We only included students if their GTAs were also participants in order to explore the correlation between the GTA instructional approach and student learning outcomes. Students who did not submit either pretest or posttest were not included.

		Physics I ($N = 425$)			Physics II ($N = 492$)		
		Semester 0 ($n = 165$)	Semester 1 ($n = 163$)	Semester 2 ($n = 97$)	Semester 0 ($n = 145$)	Semester 1 ($n = 224$)	Semester 2 ($n = 123$)
Lecture instructor	A	87 (53%)	⋯	23 (24%)	⋯	95 (42%)	⋯
	B	78 (47%)	⋯	⋯	⋯	⋯	⋯
	C	⋯	76 (47%)	⋯	⋯	⋯	⋯
	D	⋯	87 (53%)	⋯	⋯	⋯	⋯
	E	⋯	⋯	56 (58%)	⋯	⋯	⋯
	F	⋯	⋯	18 (18%)	⋯	⋯	⋯
	G	⋯	⋯	⋯	34 (23%)	129 (58%)	80 (65%)
	H	⋯	⋯	⋯	111 (77%)	⋯	43 (35%)
GTA regular approach	Interactive	110 (67%)	122 (75%)	97 (100%)	98 (68%)	224 (100%)	90 (73%)
	Not interactive	55 (33%)	41 (25%)	0 (0%)	47 (32%)	0 (0%)	33 (27%)

based on the dominant instructional style GTAs were observed to use. Controlling for student incoming preparation and lecture instruction, we used ANCOVA (or ANOVA) to evaluate the correlation between student learning and GTA instructional approach.

We evaluated data from both courses as to whether the assumptions of linear modeling are met. The data from Physics II showed homogeneity of variances, but physics I data displayed heteroskedasticity. To account for heteroskedasticity, we used the *Anova* function in the *car* package [84] with the classical White correction. While the residuals of the models were not normal (as determined by Shapiro-Wilk test [85]), the F test is considered robust to nonnormality in terms of type I error [86]. We also conducted a Kruskal-Wallis test for each parametric test, and the results were consistent.

B. Correlation between undergraduate student learning and simulator training

The number of students who participated in each semester with each lecture instructor is shown in Table VII. The participation rate associated with each lecture instructor ranged from approximately 10% to 50% (rounded to the nearest 5%), and the participation rate associated with each GTA ranged from approximately 20% to 90%[11] (rounded to the nearest 5%). Figure 10 shows the boxplots of student scores in different semesters when GTAs participated in different numbers of simulator sessions. In order to control for student incoming preparation as well as lecture instruction, we proposed a linear model in which the pretest score, lecture instructor, and semester are the independent variables, and the posttest score is the outcome. However, the lecture instructors for Physics I were found to be highly correlated with semesters (i.e., instructors were varied in each of the semesters) as shown in Table VII. Therefore, the model we used for data from Physics I did not include lecture instructor as an independent variable. Since faculty members' teaching assignments were beyond the control of the researchers, we were not able to disentangle the effect of lecture instruction in Physics I from the effect of simulator training GTAs participated in. An ANCOVA analysis (see Appendix D, Table XII) suggests that students in Physics I from different semesters had equivalent performance on the FCI posttest [$F(2, 421) = 0.2$, $p = 0.831$, $\eta^2_{\text{partial}} = 8.6 \times 10^{-4}$] when the pretest scores were controlled for.

For data from Physics II, we removed the pretest score from the model because the posttest scores were not found to be linearly dependent on pretest scores, which may be due to that the pretest scores were subject to a "floor effect" (i.e., all students scored poorly on the pretest since they were unfamiliar with the topics in electricity and magnetism [87]). Therefore, we first compared student pretest scores between semesters using one-way ANOVA. The results suggest that students in Physics II from different semesters had equivalent preparations [$F(2, 489) = 2.9$, $p = 0.054$, $\eta^2_{\text{partial}} = 0.012$].[12] We then compared student performance on posttest controlling for lecture instructors

[11]In semester 2, 7 out of 32 sections' worth of data appeared missing. The 7 sections were associated with 5 GTAs (1 or 2 per GTA). It was unclear whether those data were lost or the data associated with the same GTA in difference sections were mixed together. The missing sections were excluded from the analysis.

[12]The p value is very close to the critical value of 0.05. To avoid type II error, we conducted pairwise t tests. The results suggest that the difference between any two groups was not statistically significant.

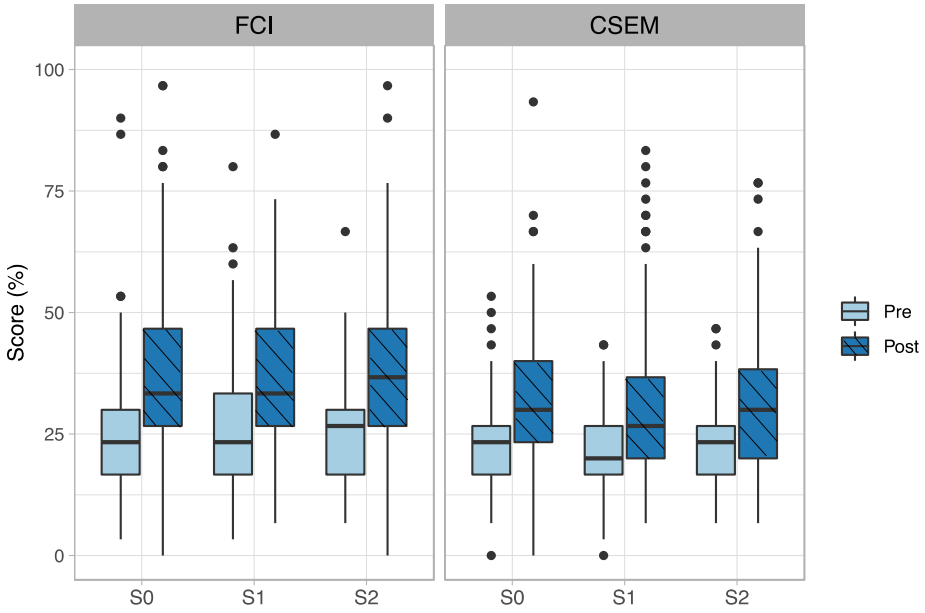

FIG. 10. Undergraduate student pretest and posttest scores on FCI and CSEM across semesters. There is no difference in student learning as measured by FCI and CSEM across semesters.

using two-way ANOVA (see Appendix D, Table XIII). The difference in student performance on posttest from different semesters was not statistically significant [$F(2, 487) = 0.1$, $p = 0.891$, $\eta^2_{\text{partial}} = 4.7 \times 10^{-4}$]. Since we did not find a difference in student posttest scores across semesters in both courses (with and without controlling for lecture instructors), we conclude that the simulator training GTAs participated in did not have a measurable impact on student conceptual understanding as measured by common concept inventories.

C. Correlation between GTA regular instructional approach and undergraduate student learning

We investigated the correlation between GTA regular instructional approach and student learning. Since some GTAs used more than one instructional style, we needed to determine the more prevalent approach each GTA used as the proxy for GTA regular approach. Considering that we only observed each GTA 3 to 5 times in each semester, many GTAs would not have a more prevalent approach if we categorize GTAs' approach using the three instructional styles identified from the cluster analysis. Therefore, we fit the three instructional styles into two major categories based on the extent to which GTAs were being interactive. The group-work facilitator and the whole-class facilitator were considered interactive while the waiter was considered not-interactive. Such grouping is reasonable because both the group-work and the whole-class facilitators had drastically higher fractions of 1o1-Talk, 1o1-TPQ, and VM compared to the waiter; they also had much lower fractions of M and W compared to the waiters. We then identified the more prevalent approach (either interactive or not interactive) for each GTA in each semester. Examining the difference in numbers of observations with different approaches, we found that in all semesters, about 73% of the GTAs had a difference of either 3 or 4. We refer to this more prevalent[13] approach as the GTA's regular instructional approach. The boxplots of student performance corresponding to each GTA regular approach across all the semesters are shown in Fig. 11.

For data from Physics I, we conducted an ANCOVA (see Appendix D, Table XIV) analysis to investigate the correlation between GTA regular instructional approach and student performance on FCI posttest, while controlling for student FCI pretest score and lecture instructor. The difference in FCI posttest scores between groups of students who were taught with different GTA regular approaches was statistically significant, with a small effect size [$F(1, 417) = 4.3$, $p = 0.040$, $\eta^2_{\text{partial}} = 0.011$].

For data from Physics II, we first performed a t test on student CSEM pretest scores between the two groups. The difference was not statistically significant [$t(490) = -0.2$, $p = 0.810$]. We then conducted a two-way ANOVA for posttest scores between the groups while controlling for lecture instructor (see Appendix D, Table XV). The difference between the groups were not statistically significant. [$F(1, 488) = 1.1$, $p = 0.295$, $\eta^2_{\text{partial}} = 2.2 \times 10^{-3}$].

[13]Only one GTA had an equal proportion of interactive and not-interactive approach. In addition, this GTA was only observed 2 times due to a change in GTA teaching assignments. Therefore, we eliminated the data (including GTA regular approach and student scores) associated with this GTA in the analysis.

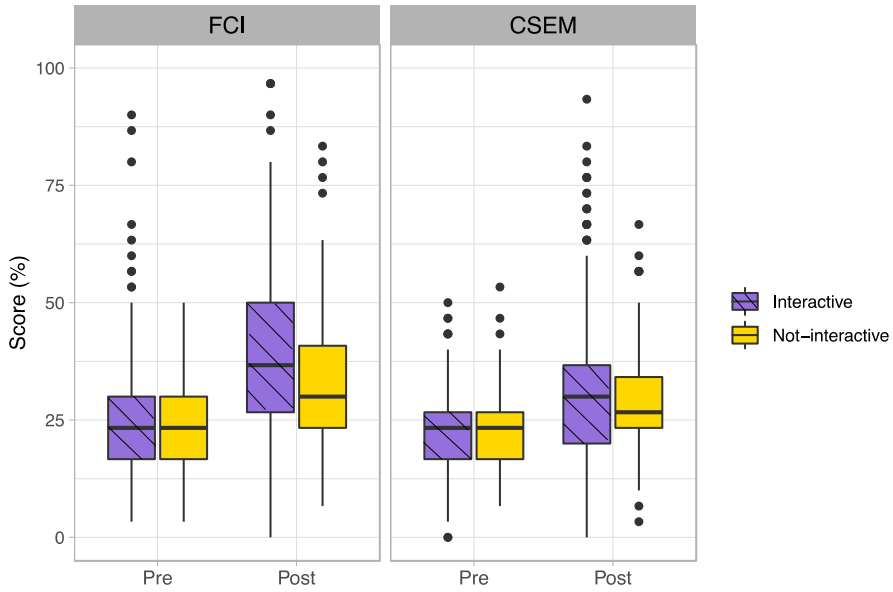

FIG. 11. Pretest and posttest scores on FCI and CSEM for undergraduate students who had GTAs with different regular instructional approaches. Student performance on FCI posttest is correlated with GTA regular instructional approach; while there is no correlation between student performance on CSEM posttest and GTA regular instructional approach.

The results from Physics I suggest that GTAs' regular instructional approach may have an impact on undergraduate student learning measured by the FCI. This is consistent with prior research that GTAs' content knowledge and teaching skills can lead to improved undergraduate performance [3–9]. However, we did not find a consistent result in Physics II where student learning was measured by CSEM. It is unclear what led to the difference in findings. We discuss possible factors in Sec. VI.

VI. DISCUSSION AND CONCLUSION

In this study, we evaluated the impact of a GTA PD program that makes use of a classroom simulator. During PD, GTAs were given opportunities to deliberately practice specific pedagogical skills in simulated classroom environments and receive feedback from PD facilitators. We examined the impact of simulator training on GTAs' classroom practices as well as the subsequent influence on undergraduate student learning. We observed GTAs teaching mini-studios using a modified version of LOPUS with additional codes for cold calling. GTAs' classroom practices were examined in two respects: (i) GTAs' instructional styles determined by a cluster analysis on the behaviors described in the modified version of LOPUS, and (ii) GTAs' implementations of questioning-related pedagogical skills. The first analysis provides a holistic means to measure instructional practices, and the second examines a specific aspect, questioning, which is an essential instructional strategy to engage students in active learning. Undergraduate student learning was measured by FCI (in Physics I) and CSEM (in Physics II). As part of the design of this study, we varied the number of simulator sessions GTAs participated in across three semesters: no simulator, one session, and four sessions. This allowed us to examine whether simulator training has an impact, and how many sessions in a semester are needed to observe an impact.

A. Deliberate practice in the simulator supports GTAs to implement interactive instructional styles and questioning-related skills in real classrooms

We identified three instructional styles that GTAs used: the group-work facilitator, the waiter, and the whole-class facilitator. Ideally, we expect GTA PD to be able to shift GTAs away from the waiters as the waiters tend to wait for students to call on them and they had fewer interactions with students. The results suggest that one-session simulator training did not seem to impact GTAs' use of instructional styles. However, four-session simulator training appeared to shift some GTAs away from the group-work facilitators and the waiters toward the whole-class facilitators. It was encouraging that the four-session simulator training was able to shift some GTAs away from the waiters, who tended to wait until students called on them and rarely asked students to answer questions. This suggests that GTAs became more interactive after the four-session simulator training. However, the fact that some GTAs shifted away from the group-work facilitators indicates an area for improvement in our PD.

During the first simulator session, GTAs were prompted to hold a whole-class discussion and to cold call. In retrospect, it was not clear whether GTAs perceived that

a whole-class discussion was also expected for two of the remaining sessions (the third session was explicitly described as checking in on individual groups). The simulator environment was limited to five students (in two groups), which made it easier to teach in a whole class setting compared to a real classroom. We also note that the head GTA in semester 2 was observed to consistently use the whole-class facilitator in all three observations. The head GTA's preference in instructional styles may have influenced how he facilitated the weekly prep meetings, which may have in turn have affected other GTAs' classroom practices. This is consistent with the framework from Reeves *et al.* [38] that implementation variables can moderate the relationship between GTA PD design and the outcome variables. The head GTA's teaching skills and beliefs about teaching can influence the implementation of GTA PD, which can in turn affect the outcomes.

It is worthwhile to mention that in every fall semester (e.g., semester 0 and semester 2), approximately half of the GTAs were new (i.e., never taught mini-studio or lab before), and typically these GTAs continued to teach the same course in the following spring semester (e.g., semester 1). Therefore, GTAs in the spring usually have more teaching experience than GTAs in the fall. Indeed, out of 10 participating GTAs in semester 1, nine GTAs had taught either in semester 0 or prior to the study. However, we did not see a difference in GTAs' use of instructional styles between semester 1 and semester 0. This suggests that GTAs do not tend to automatically change instructional styles as they gain more teaching experience. Moreover, we have seen results indicating that GTAs tend to become less interactive as they gain more teaching experience [67,88]. Therefore, GTA PD that focuses on deliberate practice is necessary to support positive change in GTAs' classroom practices.

Besides investigating GTAs' instructional practices holistically, we also examined a specific aspect, GTAs' implementations of questioning-related skills, which included cold calling, posing questions with whole class listening, and posing questions to individuals or groups of students. We found that during semester 2 when GTAs participated in four sessions of simulator training, GTAs implemented significantly higher frequencies of cold calling (a medium effect size) and posing questions with the whole class listening (a large effect size). Higher frequencies of these two skills appear to be consistent with the fact that more GTAs shifted towards the whole-class facilitators in semester 2. We did not see the same effect during semester 1 when GTAs only participated in one session of simulator training. This may not be surprising since developing PCK and teaching skills requires repeated practice and tailored feedback. The results suggest four-session simulator training supports GTAs' deliberate practice and transfer of teaching skills to actual classrooms. This finding is consistent with prior studies in K-12 contexts that science teachers who participated in four 10-min sessions were found to increase target teaching practices in both the simulator and their real classrooms [33].

B. Simulator training supports both new and more experienced GTAs who initially demonstrate differences in classroom practices

We found that new GTAs were more interactive compared to more experienced GTAs before the simulator was incorporated in PD. New GTAs made use of interactive instructional styles (the group-work facilitator and the whole-class facilitator) more often; they also posed more questions in both small-group and whole-class settings. The results also showed that four-session simulator training supported both new and more experienced GTAs to implement interactive instructional styles as well as questioning-related skills. In addition, new GTAs appeared to be somewhat more receptive to simulator training. In semester 2, none of the new GTAs made use of the style of waiter in our observations, while four class periods taught by two different experienced GTAs were in the cluster associated with the waiter.

Our results seem somewhat contradictory to a finding from French and Russell [89] in an inquiry lab section of a biology course. In their study, the postsemester survey showed that more experienced GTAs favored inquiry-style lab over traditional or a combination of inquiry and traditional labs, while new GTAs did not show a preference. However, French and Russell did not directly measure GTAs' classroom practices. We argue that both the study by French and Russell and our study reflect that GTAs' teaching beliefs and practices are influenced by the department and university culture. In French and Russell's study, GTAs gradually aligned their perception about the benefits of inquiry lab with the course developers as they gained more experience with inquiry lab. In our study, GTAs became less interactive in classrooms as they learned more about the department and/or university culture that research is highly prioritized over teaching [90,91]. Yet, the simulator training showed effects on both new and more experienced GTAs. It is likely that the GTAs received signals that the department valued teaching since many resources and efforts were allocated to GTA PD. This conjecture is consistent with prior research on GTA PD that departmental climate potentially has a strong influence on GTAs' teaching practices [92,93].

C. GTAs need more support to effectively mitigate student resistance

Although simulator training was effective at supporting GTAs to implement questioning-related skills, the frequencies for some of the target skills decreased over the course of a semester. Posing questions with whole class listening decreased in both semesters 1 and 2; posing questions to an

individual or group of students decreased in semester 1. This is consistent with the study by Becker *et al.* [28] on the effects of practice-based training on GTA classroom practices. Becker *et al.* also found that some of the questioning-related skills (cold call and stretch it: explain logic) GTAs adopted had decreased frequencies throughout the semester. Similarly, Wheeler, Maeng, and Whitworth [94] found that some TAs whose teaching beliefs shifted toward "TA as facilitator" during TA PD reverted back to "TA as disseminator" after teaching. In their study, one TA stated that this was because "he just felt he needed to tell students the answer sometimes" [94]. Indeed, GTAs' teaching beliefs as well as behaviors can be influenced by students' behaviors through GTA-student interactions [10,95]. Wilcox, Yang, and Chini found that GTAs' behaviors (in the same courses studied here) are influenced by their perceptions of student expectations [10]. Sandi-Urena, Cooper, and Gatlin reported that GTAs perceived that a main source of student frustration was that students' expectations differed substantially from what they encountered in lab [96]. In the study by Roehrig *et al.* [95], GTAs stated that "I do not want to get bad student evaluations as a result of their frustrations with inquiry-based laboratories." These studies suggest that GTAs recognize that students can feel anxious and frustrated in active learning environments, but they lack support for effectively mitigating student resistance.

It is also worthwhile to point out that there was a large variation in the extent to which lecture instructors made use of active learning strategies in the associated lecture sections. As a result, some GTAs needed to work against the norms set by the lecture instructors. It could be particularly challenging for students who had didactic lecture instruction to buy in to the active learning strategies implemented by GTAs in mini-studios. We suggest that GTA PD programs provide opportunities for GTAs to discuss issues concerning student resistance in active learning environments. In addition, GTAs should be introduced research-based strategies to mitigate student resistance (e.g., Ref. [97]).

D. Simulator training has a potential to impact undergraduate student learning

The simulator training GTAs participated in did not seem to impact undergraduate student performance on FCI and CSEM. The differences in student performance on posttests between different semesters were not statistically significant, when controlling for pretest scores and lecture instruction. However, GTAs' regular instructional approach was found to be correlated with undergraduate student performance on FCI posttest with a small effect size when both pretest scores and lecture instruction were controlled for. This is consistent with prior research that GTAs' teaching skills are linked with undergraduate student learning [3–9]. In contrast, we did not find similar results from student performance on CSEM. It is not clear whether the inconsistent results were due to low participation rate among the undergraduate students, which could have led to an unrepresentative sample. It is worthwhile to mention that students who took CSEM had also taken FCI in the previous semester. It is likely that students' efforts on CSEM decreased because neither concept inventories counted toward students' course grades. This speculation is supported by prior research that both time and incentives can influence test outcomes [98]. This may be a reason why students' posttest scores on CSEM were not linearly dependent on pretest scores. Furthermore, both FCI and CSEM only focus on a narrow scope of the content covered in introductory physics courses. Therefore, these two instruments are limited in measuring the full scope of student improvement in learning that may have resulted from GTAs' participation in the simulator training. Future research should measure student learning through multiple means (e.g., exam, homework, concept inventories) and synthesize findings.

Although we did not find an impact of simulator training on undergraduate conceptual learning, evidence showed that the simulator training has the potential to increase student learning outcomes. First, simulator training supports GTA classroom practice; GTAs made use of more interactive instructional styles and implemented more questioning-related skills. Second, some of the results showed that GTA classroom practice is correlated with undergraduate student learning. Therefore, if the simulator training further improved GTA classroom practice, a measurable impact on student learning should be found. Third, student engagement indicated by student questions increased as GTAs made use of more interactive instructional styles. Increased student engagement may in turn lead to increased learning outcomes, as prior research has demonstrated that student engagement can predict student learning in labs [7].

VII. LIMITATIONS

We encountered several logistical challenges during this study. We were not able to observe each GTA in semester 0 during the same weeks; thus, the curriculum units were not controlled for in the baseline semester, which may have resulted in variations of GTA and student behaviors. In addition, the analysis was somewhat constrained by the limited number of observations. We examined the classroom practices of GTAs in Physics I and Physics II as a whole rather than at the individual course level. We acknowledge that it is possible that the findings may differ if the data were disaggregated. Future research should explore differences in GTAs' classroom practices in different courses. Moreover, we only examined descriptively how simulator training impacted the use of instructional

styles for GTAs with and without prior teaching experience; the way we defined GTA regular instructional approach (i.e., most frequent approach among our observations of each GTA) was a little crude. Furthermore, the participation rate of undergraduate students was low. We compared the scores of the research participants to the department-level data, which included all the students who submitted both pretest and posttest (about two-thirds of total enrollment). The average scores of the research participants were slightly higher than the department-level data. However, it is likely that students who submitted both pretest and posttest had higher performance than those who did not. Therefore, our sample may be representative even if it is a small fraction of total enrollment.

Other limitations concern the scope of this paper. We did not report GTA implementation of other pedagogical skills that they rehearsed in the simulator training. We also did not explore how feedback GTAs received during the simulator training supported deliberate practice. In addition, student learning was only measured by concept inventories, which omitted impact on other aspects of learning. Finally, our study only explores the impact of simulator training at one university and does not examine the impact of simulator training across GTA identities, such as the country where GTAs completed their own undergraduate studies.

VIII. IMPLICATIONS AND FUTURE WORK

The results show that integrating simulator training in GTA PD is effective at supporting GTAs to deliberately practice evidence-based teaching skills in the simulator and to transfer these skills to real classrooms. We also found that, without the simulator training, GTAs who have more experience tend to behave less interactively compared to new GTAs. Yet, deliberate practice in the simulator has a positive impact on both new and experienced GTAs. These findings suggest that effective GTA PD is necessary to support GTAs to implement evidence-based teaching practices. We suggest GTA PD programs provide multiple opportunities for GTAs to engage in deliberate practice. To gain insights into GTA experience of deliberate practice in the simulator, we conducted interviews with GTAs. Future research will explore GTAs' perspectives of how simulator training impacts their classroom practices.

The findings in the study also indicate areas for improvement. GTAs implementations of questioning-related skills tend to decrease over the course of a semester, which may be due to student resistance to active learning environments. We suggest GTA PD programs specifically address strategies to mitigate student resistance. Future research will explore how to effectively implement this in GTA PD. We also plan to evaluate how GTAs implement other teaching skills, such as stretch-it and group facilitation. In addition, we will examine the effects of the simulator training on GTAs' classroom practices in different disciplines in order to generalize the findings (see Ref. [62] for initial findings comparing GTAs in physics and chemistry at the same institution).

ACKNOWLEDGMENTS

This work is funded in part by NSF Grant No. DUE 1725554. We acknowledge Isadore Nottolini for assistance developing materials for the simulator training. We also acknowledge Andrew Cheshire for helping organize the observation data for analysis. Last but not least, we thank the faculty members, graduate teaching assistants and undergraduate students who participated in this study.

APPENDIX A: COMPARING RESEARCH SAMPLE TO DEPARTMENT DATA

Table VIII shows the average scores with standard errors on FCI and CSEM before and after instructions from research participants and department-level data. We conducted a Welch's t test [99] for each result since Welch's t does not assume equal variance. We found that the difference is significant only on the CSEM posttest. We then calculated the effect size using Cohen's d [100], but the result ($d = 0.12$) does not meet the cutoff for a small effect ($d > 0.2$). Therefore, we concluded that the performance of the research participants was comparable to the department-level data.

TABLE VIII. The average scores with standard errors on concept inventories from research participants and department data. The p value from a Welch's t test is also shown. $^*p < 0.05$.

	FCI pretest	FCI posttest	CSEM pretest	CSEM posttest
Research participants	$25.1\% \pm 0.6\%$ $N = 425$	$37.3\% \pm 0.8\%$ $N = 425$	$23.0\% \pm 0.4\%$ $N = 492$	$30.8\% \pm 0.7\%$ $N = 492$
Department data	$24.6\% \pm 0.4\%$ $N = 1201$	$36.1\% \pm 0.5\%$ $N = 1201$	$22.6\% \pm 0.2\%$ $N = 1061$	$29.1\% \pm 0.4\%$ $N = 1061$
p (Welch's t)	0.488	0.204	0.371	0.035*

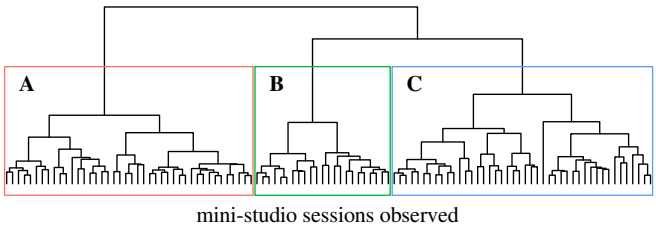

FIG. 12. Dendrogram showing the hierarchical structure of the mini-studio sessions observed. Clusters A, B, and C are indicated in the boxes.

APPENDIX B: CLUSTER ANALYSIS

We conducted a cluster analysis with 109 mini-studio sessions observed and found three clusters. The cluster structure is shown in Fig. 12. To investigate how each cluster differs from one another, we conducted a Kruskal-Wallis test, as well as a Dunn's test with Holm-Bonferroni correction, for each code. The results are shown in Table IX.

APPENDIX C: DIFFERENCE IN PROPORTIONS OF GTAS' USE OF INSTRUCTIONAL STYLES BETWEEN PHYSICS I AND PHYSICS II

We examined the representation of Physics I and Physics II mini-studio class periods in different clusters. Table X shows the number of class periods observed in different clusters in each course. The chi-squared test suggested that the difference in proportions of class periods between the two courses is statistically significant with a medium effect size $[\chi^2(2) = 7.2,\ p = 0.027,$ Cramer's $V = 0.238]$. However, multiple pairwise comparisons with Holm-Bonferroni correction did not show a significant difference between any pair of clusters. The adjusted p values are

TABLE X. Distributions of class periods observed in different clusters in each course.

	Cluster A: The group-work facilitators	Cluster B: The waiters	Cluster C: The whole-class facilitators
Physics I ($n = 59$)	16	14	29
Physics II ($n = 50$)	26	9	15

TABLE XI. Participation rates for GTAs in different courses in each semester.

	Physics I GTAs	Physics II GTAs
Semester 0	4/7 (57%)	4/6 (67%)
Semester 1	5/7 (71%)	5/6 (83%)
Semester 2	10/12 (83%)	5/6 (83%)

TABLE IX. Results from Kruskal-Wallis and Dunn's tests with Holm-Bonferroni correction. $*p < 0.05$. $**p < 0.01$. $***p < 0.001$. † small effect size. †† medium effect size. ††† large effect size. The p values from Dunn's tests are bolded if the difference is significant at the 0.05 level with Holm-Bonferroni correction.

				Pairwise p		
Code	$\chi^2(2)$	p	η^2	A vs B	A vs C	B vs C
Lec	44.3	2.439×10^{-10}***	0.399†††	**0.005**	<**0.001**	**0.003**
RtW	14.6	6.849×10^{-4}***	0.119††	0.056	<**0.001**	0.110
FUp	10.8	0.004**	0.083††	0.250	**0.009**	**0.006**
D/V	21.2	2.463×10^{-5}***	0.181†††	0.101	<**0.001**	<**0.001**
M	35.1	2.349×10^{-8}***	0.313†††	<**0.001**	**0.004**	<**0.001**
1o1-Talk	55.5	9.025×10^{-13}***	0.504†††	<**0.001**	<**0.001**	<**0.001**
PQ	31.8	1.229×10^{-7}***	0.281†††	0.335	<**0.001**	<**0.001**
1o1-TPQ	37.7	6.452×10^{-9}***	0.337†††	<**0.001**	**0.013**	<**0.001**
VF	18.6	9.01×10^{-5}***	0.157†††	<**0.001**	0.265	<**0.001**
VM	45.5	$1. \times 10^{-10}$ ***	0.411†††	<**0.001**	**0.001**	<**0.001**
TI	20.2	4.146×10^{-5}***	0.172†††	<**0.001**	**0.013**	**0.008**
Adm	2.6	0.272	0.006	0.395	0.164	0.313
W	29.9	3.285×10^{-7}***	0.263†††	<**0.001**	**0.002**	**0.002**
Wks/Lab	34.9	2.644×10^{-8}***	0.310†††	0.315	<**0.001**	<**0.001**
TQ	6.4	0.040*	0.042†	0.027	0.069	0.193
SQ	47.8	4.125×10^{-11}***	0.432†††	**0.001**	<**0.001**	**0.005**
1o1-SQ	9.5	0.009**	0.071††	**0.003**	0.047	0.094
WC	2.9	0.237	0.008	0.388	0.137	0.293
SI	1.1	0.583	−0.009	0.531	0.413	0.402

$p_{adj,A\text{-}B} = 0.399$, $p_{adj,A\text{-}C} = 0.054$, and $p_{adj,B\text{-}C} = 1.000$, respectively.

We speculate that our samples of GTAs in Physics I and Physics II may not be representative to allow for a comparison between the two courses. As shown in Table XI, the samples of GTA in different courses in each semester are very small. In addition, Physics I GTAs had lower participation rates than Physics II GTAs in both semester 0 and semester 1. Furthermore, the number of GTAs from Physic I was twice the number of GTAs from Physics II due to increased student enrollment in Physics I. We argue that larger samples are needed to evaluate whether GTAs in Physics I and Physics II are different in their classroom practices, and whether the simulator training is effective in individual courses as opposed to being effective in the introductory sequence as a whole.

APPENDIX D: COMPLETE RESULTS FROM MULTIPLE ANCOVA AND ANOVA ANALYSES

TABLE XII. Results of one-way ANCOVA (type III) conducted on student performance on FCI posttest in Physics I with semester as an independent variable and pretest score as a covariate. Heteroscedasticity correction was conducted.

Independent variable	df	F	p	$\eta^2_{partial}$
Intercept	1	103.5	$<2.2 \times 10^{-16}$***	
Pretest score	1	67.3	2.896×10^{-15}***	0.23
Semester	2	0.2	0.831	8.6×10^{-4}
Residuals	421			

TABLE XIII. Results of two-way ANOVA (type III) conducted on student performance on CSEM posttest in Physics II with semester and lecture instructor as independent variables.

Independent variable	df	F	p	$\eta^2_{partial}$
Intercept	1	557.0	$<2.2 \times 10^{-16}$***	
Semester	2	0.1	0.891	4.7×10^{-4}
Lecture instructor	2	7.4	7.127×10^{-4}***	0.029
Residuals	487			

TABLE XIV. Results of two-way ANCOVA (type III) conducted on student performance on FCI posttest in Physics I with lecture instructor and GTA regular approach as independent variables and pretest score as a covariate. Heteroscedasticity correction was conducted.

Independent variable	df	F	p	$\eta^2_{partial}$
Intercept	1	58.0	1.773×10^{-13}***	
Pretest score	1	73.8	$<2.2 \times 10^{-16}$***	0.23
GTA regular instructional approach	1	4.3	0.040*	0.011
Lecture instructor	5	4.1	0.001**	0.044
Residuals	417			

TABLE XV. Results of two-way ANOVA (type III) conducted on student performance on CSEM posttest in Physics II with GTA regular approach and lecture instructor as independent variables.

Independent variable	df	F	p	$\eta^2_{partial}$
Intercept	1	483.0	$<2.2 \times 10^{-16}$***	
GTA regular instructional approach	1	1.3	0.259	2.2×10^{-3}
Lecture instructor	2	8.4	2.546×10^{-4}***	0.031
Residuals	488			

[1] E. Seymour, *Partners in Innovation: Teaching Assistants in College Science Courses* (Rowman & Littlefield, Washington, DC, 2005).

[2] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth, Active learning increases student performance in science, engineering and mathematics, Proc. Natl. Acad. Sci. U.S.A. **111**, 8410 (2014).

[3] T. J. Greenbowe and B. Hand, Introduction to the science writing heuristic, in *Chemists' Guide to Effective Teaching*, 1st ed. (Prentice Hall Englewood Cliffs, NJ, 2005).

[4] Z. Hazari, A. W. Key, and J. Pitre, Interactive and affective behaviors of teaching assistants in a first year physics laboratory, Electron. J. Res. Sci. Math. Educ. **7**, 1 (2003).

[5] K. M. Koenig, R. J. Endorf, and G. A. Braun, Effectiveness of different tutorial recitation teaching methods and its implications for TA training, Phys. Rev. ST Phys. Educ. Res. **3**, 010104 (2007).

[6] J. R. Poock, K. A. Burke, T. J. Greenbowe, and B. M. Hand, Using the science writing heuristic in the general chemistry laboratory to improve students' academic performance, J. Chem. Educ. **84**, 1371 (2007).

[7] J. B. Stang and I. Roll, Interactions between teaching assistants and students boost engagement in physics labs, Phys. Rev. ST Phys. Educ. Res. **10**, 020117 (2014).

[8] L. B. Wheeler, J. L. Maeng, J. L. Chiu, and R. L. Bell, Do teaching assistants matter? Investigating relationships between teaching assistants and student outcomes in undergraduate science laboratory classes, J. Res. Sci. Teach. **54**, 463 (2017).

[9] A. S. Huffmyer and J. D. Lemus, Graduate TA teaching behaviors impact student achievement in a research-based undergraduate science course, J. Coll. Sci. Teach. **048**, 56 (2019).

[10] M. Wilcox, Y. Yang, and J. J. Chini, Quicker method for assessing influences on teaching assistant buy-in and practices in reformed courses, Phys. Rev. Phys. Educ. Res. **12**, 020123 (2016).

[11] J. S. Boman, Graduate student teaching development: evaluating the effectiveness of training in relation to graduate student characteristics, Can. J. High. Educ. **43**, 100 (2013).

[12] E. M. Duffy and M. M. Cooper, Assessing TA buy-in to expectations and alignment of actual teaching practices in a transformed general chemistry laboratory course, Chem. Educ. Res. Pract. **21**, 189 (2020).

[13] J. A. Luft, J. P. Kurdziel, G. H. Roehrig, and J. Turner, Growing a garden without water: Graduate teaching assistants in introductory science laboratories at a doctoral/research university, J. Res. Sci. Teach. **41**, 211 (2004).

[14] J. M. Mutambuki and R. Schwartz, We don't get any training: the impact of a professional development model on teaching practices of chemistry and biology graduate teaching assistants, Chem. Educ. Res. Pract. **19**, 106 (2018).

[15] R. A. B. Rodriques and J. Bond-Robinson, Comparing faculty and student perspectives of graduate teaching assistants' teaching, J. Chem. Educ. **83**, 305 (2006).

[16] E. A. West, C. A. Paul, D. Webb, and W. H. Potter, Variation of instructor-student interactions in an introductory interactive physics course, Phys. Rev. ST Phys. Educ. Res. **9**, 010109 (2013).

[17] L. R. Prieto and K. R. Scheel, Teaching assistant training in counseling psychology, Counseling Psychol. Quart. **21**, 49 (2008).

[18] B. T. Spike and N. D. Finkelstein, Preparing tutorial and recitation instructors: A pedagogical approach to focusing attention on content and student reasoning, Am. J. Phys. **80**, 1020 (2012).

[19] J. R. Thompson, W. M. Christensen, and M. C. Wittmann, Preparing future teachers to anticipate student difficulties in physics in a graduate-level course in physics, pedagogy, and education research, Phys. Rev. ST Phys. Educ. Res. **7**, 010108 (2011).

[20] V. Otero, S. Pollock, and N. Finkelstein, A physics department's role in preparing physics teachers: The Colorado learning assistant model, Am. J. Phys. **78**, 1218 (2010).

[21] K. A. Ericsson, R. T. Krampe, and C. Tesch-Romer, The role of deliberate practice in the acquisition of expert performance, Psychol. Rev. **100**, 363 (1993).

[22] L. Darling-Hammond, Constructing 21st-century teacher education, J. Teach. Educ. **57**, 300 (2006).

[23] K. Zeichner, The turn once again toward practice-based teacher education, J. Teach. Educ. **63**, 376 (2012).

[24] M. S. Garet, A. C. Porter, L. Desimone, B. F. Birman, and K. S. Yoon, What makes professional development effective? Results from a national sample of teachers, Am. Educ. Res. J. **38**, 915 (2001).

[25] M. Lampert, H. Beasley, H. Ghousseini, E. Kazemi, and M. Franke, Using designed instructional activities to enable notices to manage ambitious mathematics teaching, in *Instructional Explanations in the Disciplines* (Springer, New York, 2010), pp. 129–141.

[26] A. Remesh, Microteaching, an efficient technique for learning effective teaching, J. Res. Med. Sci. **18**, 158 (2013).

[27] E. Etkina, Pedagogical content knowledge and preparation of high school physics teachers, Phys. Rev. ST Phys. Educ. Res. **6**, 020110 (2010).

[28] E. A. Becker, E. J. Easlon, S. C. Potter, A. Guzman-Alvarez, J. M. Spear, M. T. Facciotti, M. M. Igo, M. Singer, and C. Pagliarulo, The effects of practice-based training on graduate teaching assistants' classroom practices, CBE Life Sci. Educ. **16**, ar58 (2017).

[29] M. Dawson and B. Lignugaris/Kraft, TeachLivE vs. roleplay: Comparative effects on special educators acquisition of basic teaching skills, in *Conference Proceedings for First National TLE TeachLivE Conference, Orlando, FL, May 23–24, 2013*, edited by A. Hayes, S. Hardin, L. Dieker, M. Hynes, C. Hughes, and C. Straub (University of Central Florida, Orlando, FL, 2013).

[30] A. H. Brown, Simulated classrooms and artificial students: The potential effects of new technologies on teacher education, J. Res. Comput. Educ. **32**, 307 (1999).

[31] E. Hixon and H. So, Technology's role in field experiences for preservice teacher training, Educ. Technol. Soc. **12**, 294 (2009).

[32] K. Vanlehn, S. Ohlsson, and R. Nason, Applications of simulated students: An exploration, J. Artificial Intelligence Educ. **5**, 135 (1994).

[33] C. Straub, L. Dieker, M. Hynes, and C. Hughes, Using virtual rehearsal in the TLE TeachLivE™ mixed reality classroom simulator to determine the effects on the of performance of science teachers: A follow-up study (year 2), in *2015 TeachLivE National Research Project: Year 2 Findings* (University of Central Florida, Orlando, FL, 2015).

[34] K. V. Garland, E. Vasquez III, and C. Pearl, Efficacy of individualized clinical coaching in a virtual reality classroom for increasing teachers' fidelity of implementation of discrete trial teaching, Educ. Training Autism and Develop. Disabilities **47**, 502 (2012).

[35] M. Elford, S. James, and H. Haynes-Smith, Literacy instruction for pre-service educators in virtual learning environments, in *Conference Proceedings for First National TLE TeachLivE Conference., Orlando, FL, 2013*, edited by A. Hayes, S. Hardin, L. Dieker, M. Hynes, C. Hughes, and C. Straub (University of Central Florida, Orlando, FL, 2013).

[36] E. Whitten, A. Enicks, L. Wallace, and D. Morgan, Study of a mixed reality virtual environment used to increase teacher effectiveness in a pre-service preparation program, in *Conference Proceedings for First National TLE TeachLivE Conference, Orlando, FL, 2013*, edited by A. Hayes, S. Hardin, L. Dieker, M. Hynes, C. Hughes, and C. Straub (University of Central Florida, Orlando, FL, 2013).

[37] J. J. Chini, C. L. Straub, and K. H. Thomas, Learning from avatars: Learning assistants practice physics pedagogy in a classroom simulator, Phys. Rev. Phys. Educ. Res. **12**, 010117 (2016).

[38] T. D. Reeves, G. Marbach-Ad, K. R. Miller, J. Ridgway, G. E. Gardner, E. E. Schussler, and E. W. Wischusen, A conceptual framework for graduate teaching assistant professional development evaluation and research. CBE Life Sci. Educ. **15**, es2 (2016).

[39] T. Addy and M. R. Blanchard, The problem with reform from the bottom up: Instructional practices and teacher beliefs of graduate teaching assistants following a reformed-minded university teacher certificate programme, Int. J. Sci. Educ. **32**, 1045 (2010).

[40] J. J. Chini and J. W. T. Pond, Comparing Traditional and studio courses through FCI gains and losses, in *Proceedings of the 2014 Physics Education Research Conference, Minneapolis, MN* (AIP, New York, 2014), pp. 51–54.

[41] A. Elby, R. E. Scherr, T. McCaskey, R. Hodges, E. F. Redish, D. Hammer, and T. Bing, Open Source Tutorials in Physics Sensemaking: Suite I (2007), https://www.physport.org/curricula/MD_OST/.

[42] E. Etkina and A. Van Heuvelen, Investigative science learning environment—A science process approach to learning physics, in *PER-based reforms in calculus-based physics*, edited by E. F. Redish and P. Cooney (American Association of Physics Teachers, College Park, MD, 2007), Vol. 1, pp. 1–48.

[43] D. Lemov, *Teach Like a Champion: 49 Techniques that Put Students on the Path to College*, 1st ed. (Jossey-Bass, San Francisco, CA, 2010).

[44] S. L. Eddy, M. Converse, and M. P. Wenderoth, POR-TAAL: A classroom observation tool assessing evidence-based teaching practices for active learning in large science, technology, engineering, and mathematics classes, CBE Life Sci. Educ. **14**, ar23 (2015).

[45] E. J. Dallimore, J. H. Hertenstein, and M. B. Platt, Impact of cold-calling on student voluntary participation, J. Manage. Educ. **37**, 305 (2012).

[46] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.17.010146 for an example activity (including a physics activity, suggested student avatars' ideas, and pedagogical goals) provided to the simulator team.

[47] A. A. Geraets, I. L. Nottolini, C. M. Doty, T. Wan, J. J. Chini, and E. K. H. Saitta, Preparing GTAs for active learning in the general chemistry lab: Development of an evidence-based rehearsal module for a mixed-reality teaching simulator, J. Sci. Edu. Technol., https://doi.org/10.1007/s10956-021-09923-2 (2021).

[48] M. Broeckelman-Post, A. Johnson, and J. R. Schwebach, Calling on students using notecards: Engagement and countering communication anxiety in a large lecture, J. Coll. Sci. Teach. **045**, 27 (2016).

[49] E. J. Dallimore, J. H. Hertenstein, and M. B. Platt, Nonvoluntary class participation in graduate discussion courses: Effects of grading and cold calling, J. Manage. Educ. **30**, 354 (2006).

[50] J. K. Knight, S. B. Wise, and S. Sieke, Group random call can positively affect student in-class clicker discussions, CBE Life Sci. Educ. **15**, ar56 (2016).

[51] K. M. Cooper, V. R. Downing, and S. E. Brownell, The influence of active learning practices on student anxiety in large-enrollment college science classrooms, Int. J. STEM Educ. **5**, 23 (2018).

[52] V. R. Downing, K. M. Cooper, J. M. Cala, L. E. Gin, and S. E. Brownell, Fear of negative evaluation and student anxiety in community college active-learning science courses, CBE Life Sci. Educ. **19**, ar20 (2020).

[53] D. Watson and R. Friend, Measurement of social-evaluative anxiety, J. Consulting Clinical Psychol. **33**, 448 (1969).

[54] J. W. Weeks, R. G. Heimberg, D. M. Fresco, T. A. Hart, C. L. Turk, F. R. Schneier, and M. R. Liebowitz, Empirical validation and psychometric evaluation of the Brief Fear of Negative Evaluation Scale in patients with social anxiety disorder, Psychol. Assess. **17**, 179 (2005).

[55] B. S. Bell and S. W. Kozlowski, Active learning: Effects of core training design elements on self-regulatory processes, learning, and adaptability, 2008, from Cornell University, ILR School site http://digitalcommons.ilr.cornell.edu/articles/410.

[56] D. Steele-Johnson and Z. T. Kalinoski, Error framing effects on performance: Cognitive, motivational, and affective pathways, J. Psychol.: Interdisciplinary Applied **148**, 93 (2014).

[57] K. M. Cooper, M. Ashley, and S. E. Brownell, A bridge to active learning: A summer bridge program helps students maximize their active-learning experiences and the active-learning experiences of others, CBE Life Sci. Educ. **16** (2017).

[58] S. M. Leupen, K. L. Kephart, and L. C. Hodges, Factors influencing quality of team discussion: Discourse analysis in an undergraduate team-based learning biology course, CBE Life Sci. Educ. **19**, ar7 (2020).
[59] T. Willoughby, E. Wodd, C. McDermott, and J. McLaren, Enhancing learning through strategy instructional and group interaction: Is active generation of elaborations critical?, Appl. Cogn. Psychol. **14**, 19 (2000).
[60] J. Dunlosky, K. A. Rawson, E. J. Marsh, M. J. Nathan, and D. T. Willingham, Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology, Psychol. Sci. Publ. Interest **14**, 4 (2013).
[61] D. W. Johnson and R. T. Johnson, An educational psychology success story: Social interdependence theory and cooperative learning, Educ. Res. **38**, 365 (2009).
[62] D. W. Johnson, R. T. Johnson, and K. A. Smith, Cooperative learning: improving university instruction by basing practice on validated theory, J. Excellence University Teaching **25**, 85 (2014).
[63] L. Springer, M. E. Stanne, and S. Donovan, Measuring the success of small-group learning in college level SMET teaching: a meta-analysis, Rev. Educ. Res. **69**, 21 (1999).
[64] A. J. Lamm, C. Shoulders, T. G. Roberts, T. A. Irani, L. J. U. Snyder, and J. Brendemuhl, The influence of cognitive diversity on group problem solving strategy, J. Agricultural Educ. **53**, 18 (2012).
[65] L. D. Colin and R. E. Scherr, Making space to sensemake: Epistemic distancing in small group physics discussions, Cognit. Instr. **36**, 396 (2018).
[66] J. B. Velasco, A. Knedeisen, D. Xue, T. L. Vickrey, M. Abebe, and M. Stains, Characterizing instructional practices in the laboratory: The laboratory observation protocol for undergraduate STEM, J. Chem. Educ. **93**, 1191 (2016).
[67] T. Wan, A. A. Geraets, C. M. Doty, E. K. H. Saitta, and J. J. Chini, Characterizing science graduate teaching assistants' instructional practices in reformed laboratories and tutorials, Int. J. STEM Educ. **7**, 30 (2020).
[68] N. Wongpakaran, T. Wongpakaran, D. Wedding, and K. L. Gwet, A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples, BMC Medical Res. Methodology **13**, 61 (2013).
[69] RStudio Team, *RStudio: Integrated Development Environment for R. RStudio* (PBC, Boston, MA 2020), http://www.rstudio.com/.
[70] M. Stains *et al.*, Anatomy of STEM teaching in North American universities, Science **359**, 1468 (2018).
[71] J. H. Ward, Hierarchical grouping to optimize an objective function, J. Am. Stat. Assoc. **58**, 236 (1963).
[72] M. A. Syakur, B. K. Khotimath, E. M. S. Rochman, and B. D. Satoto, Integration K-means clustering method and elbow method for identification of the best customer profile cluster, IOP Conf. Ser.: Mater. Sci. Eng. **336**, 012017 (2018).
[73] R. Tibshirani, G. Walther, and T. Hastie, Estimating the number of clusters in a data set via the gap statistic, J. R. Stat. Soc. Ser. B **63**, 411 (2001).
[74] A. Kassambara and F. Mundt (2017). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.5, https://CRAN.R-project.org/package=factoextra.
[75] W. H. Kruskal and W. A. Wallis, Use of ranks in one-criterion variance analysis, J. Am. Stat. Assoc. **47**, 583 (1952).
[76] M. Tomczak and E. Tomczak, The need to report effect size estimates revisited. An overview of some recommended measures of effect size, Trends Sport Sci. **1**, 19 (2014).
[77] J. M. Maher, J. C. Markey, and D. Ebert-May, The other half of the story: Effect size analysis in quantitative research, CBE Life Sci Educ. **12**, 345 (2013).
[78] O. J. Dunn, Multiple comparisons using rank sums, Technometrics **6**, 241 (1964).
[79] A. Dinno, Dunn.test: Dunn's test of multiple comparisons using rank sums (2017). R package version 1.3.5. https://CRAN.R-project.org/package=dunn.test.
[80] S. Holm, A simple sequentially rejective multiple test procedure, Scand. J. Stat. Theory Appl. **6**, 65 (1979).
[81] D. C. Brooks, Space and consequences: The impact of different formal learning spaces on instructor and student behavior, J. Learning Spaces **1** (2012).
[82] Salvatore Mangiafico, rcompanion: Functions to support extension education program evaluation. R package version 2.3.25 (2020), https://CRAN.R-project.org/package=rcompanion.
[83] N. Cliff, Dominance statistics: Ordinal analyses to answer ordinal questions, Psychol. Bull. **114**, 494 (1993).
[84] J. Fox and S. Weisberg, *An {R} Companion to Applied Regression*, 3rd ed. (Sage, Thousand Oaks, CA, 2019).
[85] S. S. Shapiro and M. B. Wilk, An analysis of variance test for normality (complete samples), Biometrika **52**, 591 (1965).
[86] M. J. Blanca, R. Alarcon, J. Arnau, R. Bono, and R. Bendayan, Non-normal data: Is ANOVA still a valid option?, Psicothema **29**, 552 (2017).
[87] A. Madsen, S. B. McKagan, and E. C. Syare, Gender gap on concept inventories in. physics: What is consistent, what is inconsistent, and what factors influence the gap?, Phys. Rev. ST Phys. Educ. Res. **9**, 020121 (2013).
[88] T. Wan, C. M. Doty, A. A. Geraets, E. K. H. Saitta, and J. J. Chini, Characterizing graduate teaching assistants' teaching practices in physics "mini-studio", in *Proceedings of the 2019 Physics Education Research Conference, Provo, UT* (AIP, New York, 2019).
[89] D. French and C. Russell, Do graduate teaching assistants benefit from teaching inquiry-based laboratories?, BioScience **52**, 1036 (2002).
[90] J. D. Nyquist, L. Manning, D. H. Wulff, A. E. Austin, J. Sprague, P. K. Fraser, C. Calcagno, and B. Woodford, On the road to becoming a professor: the graduate student experience, Change **31**, 18 (1999).
[91] G. E. Gardner and M. G. Jones, Pedagogical preparation of science graduate teaching assistant: challenges and implications. Sci. Educ. **20**, 31 (2011).

[92] R. M. Goertzen, R. E. Scherr, and A. Elby, Accounting for tutorial teaching assistants' buy-in to reform instruction, Phys. Rev. ST Phys. Educ. Res. **5**, 020109 (2009).

[93] R. Sharpe, A framework for training graduate teaching assistants, Teacher Develop. **4**, 131 (2000).

[94] L. B. Wheeler, J. L. Maeng, and B. A. Whitworth, Characterizing teaching assistants' knowledge and beliefs following professional development activities within an inquiry-based general chemistry context, J. Chem. Educ. **94**, 19 (2017).

[95] G. H. Roehrig, J. A. Luft, J. P. Kurdziel, and J. A. Turner, Graduate teaching assistants and inquiry-based instruction: Implications for graduate teaching assistant training, J. Chem. Educ. **80**, 1206 (2003).

[96] S. Sandi-Urena, M. Cooper, and T. A. Gatlin, Graduate teaching assistants' epistemological and metacognitive development, Chem. Educ. Res. Pract. **12**, 92 (2011).

[97] S. Tharayil, M. Borrego, M. Prince, K. A. Nguyen, P. Shekhar, C. J. Finelli, and C. Waters, Strategies to mitigate student resistance to active learning, Int. J. STEM Educ. **5**, 7 (2018).

[98] L. Ding, N. W. Reay, A. Lee, and L. Bao, Effects of testing conditions on conceptual survey results, Phys. Rev. ST Phys. Educ. Res. **4**, 010112 (2008).

[99] B. L. Welch, The generalization of "Student's" problem when several different population variances are involved, Biometrika **34**, 28 (1947).

[100] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Routledge Academic, New York, 1988).