Impact of high-intensity training with a mixed-reality simulator on graduate teaching assistants use of questioning

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

[1]*Department of Physics, University of Central Florida, 4111 Libra Drive, Orlando, Florida 32816, USA*
[2]*Department of Chemistry, University of Central Florida, 4111 Libra Drive, Orlando, Florida 32816, USA*

(Received 30 March 2022; revised 28 March 2023; accepted 30 May 2023; published 5 July 2023)

Physics graduate teaching assistants (GTAs) are tasked with multifaceted teaching assignments, such as leading tutorials and inquiry-based laboratories, yet their professional development rarely includes opportunities to rehearse complex pedagogical skills or receive feedback on their teaching. In this study, physics GTAs practiced specific pedagogical skills during four sessions in a mixed-reality classroom simulator; here, we focus on GTAs' use of a specific questioning strategy called "Stretch-It." GTAs can apply Stretch-It by asking students to explain their logic (Explain Logic), either by explaining their work or providing evidence for their claim; or by asking students to take the content further (Follow-Up), either by applying it in an analogous situation or answering the initial question in another way. We found GTAs used all four types of Stretch-It subcategories during simulator sessions that incorporated facilitator feedback about their use of questioning. We also compared the use of Stretch-It questioning in the classroom for a pretraining semester cohort and the high-intensity simulator training cohort and found that the high-intensity cohort's average use of Explain Logic questions in the observation immediately following the Stretch-It rehearsal was meaningfully higher than the pretraining cohort's average use of Explain Logic. However, the high-intensity cohort's use of Explain Logic is unstable, and values fall back within the pretraining average in subsequent classroom observations. We did not find a significant difference between the cohorts' use of Follow-Up. We discuss the implications of our findings for science, technology, engineering, and medicine GTA professional development and the significance of providing feedback to GTAs about their teaching.

DOI: 10.1103/PhysRevPhysEducRes.19.020101

I. INTRODUCTION

Responding to national calls to transform instruction to support student learning (e.g., [1,2]), many physics departments have adopted student-centered curricula in recitations and labs often led by graduate teaching assistants (GTAs). However, such curricula require complex pedagogical skills, some of which GTAs are unfamiliar with or uncomfortable implementing. Furthermore, science, technology, engineering, and medicine (STEM) GTA training varies across universities and departments within the same university and may not include a focus on disciplinarily specific skills [3–5]. Using a mixed-reality classroom simulator, we created opportunities for GTAs to practice complex pedagogical skills and to receive immediate feedback on their use of teaching strategies within the context of the courses they teach. In this study, we investigate GTAs' use of questioning during mixed-reality simulator sessions focused on questioning and subsequent mixed-reality simulator sessions with other focal pedagogical skills as well as in their actual classroom. We introduced GTAs to a particular questioning skill called "Stretch-It" [6], which provides ideas about how to formulate questions that encourage students to explain their logic or follow-up on their understanding, as discussed in more detail below. We aim to answer the following research questions:

1. Do mixed-reality classroom simulator sessions support GTAs to rehearse Stretch-It (questioning technique)?
2. What is the impact of practice during simulator sessions on GTAs' use of Stretch-It (questioning technique) in their actual classroom?

II. BACKGROUND

A. Importance of GTA professional development

Since university physics departments often depend on GTAs to teach the recitation and laboratory sections of introductory physics courses, GTAs are crucial to postsecondary STEM education. In physics, for example, recent studies have found GTA-initiated interactions with students

*Corresponding author.
jchini@ucf.edu

in physics labs [7] and GTAs' use of questioning (instead of confirmation) during tutorials [8] are both linked with improved student performance.

However, GTAs might have a challenging time using evidence-based practices or teaching behaviors supportive of an active learning environment. In interviews with biology graduate students across the United States, Goodwin *et al.* found that while 24 out of 32 interviewees were in favor of and intended to use evidence-based teaching practices in their classrooms, only 5 out of 32 interviewees successfully implemented such practices in their classrooms [9]. One potential explanation for this finding is that GTAs do not have proper support for using evidence-based practices in the classroom.

Since physics GTAs are often new to teaching, they vary in their teaching behaviors as they develop their teaching identity. Related prior research investigating GTA teaching in tutorial and lab spaces has documented variation in teaching behaviors used by physics GTAs, including behaviors not supportive of the learning environment (i.e., waiting for students to initiate interactions) [10,11]. Similarly, education research has documented that STEM faculty may leave out important aspects of reformed instructional strategies, damping the impact of active learning on student learning [12–14]. Therefore, if we expect GTAs to implement student-centered instruction, especially if they intend to become STEM faculty, we need to create opportunities for them to develop pedagogical skills that support such instruction early in their career. Additionally, STEM GTA professional development has been found to influence GTA self-image as an instructor [15], GTA teaching confidence (correlated with GTA teaching practices [5]) [16], and GTA teaching beliefs [17,18]. Therefore, it is crucial to reform and investigate GTA training.

1. Landscape of STEM GTA training

STEM GTAs receive a variety of pedagogical preparation across institutions [3–5,18]. Typically, university-supported training is a one-time workshop at the beginning of a GTA's teaching career and often focuses on policies rather than classroom discourse or pedagogy [4,18]. STEM GTAs have described this type of training to be vague and have voiced their need for instructional support from their department [19–21]. In addition, STEM departments within the same institution likely offer various levels of training for GTAs. Examples include presemester workshops [19], semester-long pedagogy seminars [17,22], and weekly preparation meetings [3,23]. Presemester workshops might help GTAs feel more prepared to start teaching but do not provide opportunities for support during GTAs' teaching assignments [19]. Weekly preparation meetings provide continued support to GTAs during their teaching assignment, but meetings often focus on discipline-specific content or classroom management and not teaching strategies [3]. Pedagogy seminars fill in the gap left by preparation meetings by helping GTAs to develop pedagogical content knowledge (PCK) [22], but the application of PCK to classroom experiences might not be intuitive. For example, Wilcox *et al.* compared physics GTA interview responses with GTA teaching behaviors during classroom observations for GTAs who attended preparation meetings and a pedagogy seminar [24]. They found most GTAs were in favor of student-centered instruction (referred to as "buy-in"), but their teaching behaviors did not reflect their beliefs [24] This is similar to findings in biology [9]. Thus, further development of GTA training is necessary to bridge the gap between pedagogy and teaching practices.

2. Microteaching in GTA training

Gardner and Jones recommended GTA training programs should scaffold the development of PCK to encourage GTAs' use of evidence-based strategies aligned with the model of instruction and learning goals of the course [18]. One way to do this is to incorporate opportunities for GTAs to practice evidence-based strategies and receive immediate feedback (i.e., deliberate practice [25]). Microteaching is a popular training exercise implemented in preservice teacher preparation that follows a deliberate practice model. In microteaching, novice teachers take turns teaching a lesson while their peers "play" students; novice teachers then receive feedback from facilitators and their peers about their performance [26]. Research on K-12 preservice teacher training has documented reported benefits of microteaching sessions, like increased teaching self-efficacy [27], perceived preparedness to teach real students [28], and increased knowledge about reform-based teaching strategies [29]. Additionally, Fernandez and Robinson found student teachers were in favor of microteaching sessions because it allowed them to practice using pedagogical skills as one student teacher in their sample stated, "*We spend a lot of time discussing theory. It's good to get a chance to try using some of it and get feedback on how well we did before we get our own classrooms.*" [30] (p. 207).

While microteaching is not yet common in STEM GTA preparation, some benefits have been documented in STEM education literature. When surveying first-time physics GTAs about activities included in a semester-long training course, Alicea-Muñoz found GTAs valued microteaching sessions as useful to their training [31]. One GTA commented, "*I was SO nervous at first, but once we did the microteaching and labsim [another activity] and I could see what my peers were doing, I felt way better.*" [31] (p. 54). Doucette *et al.* found a positive impact of incorporating "role-playing" sessions into weekly preparation meetings on physics GTA-student interactions (e.g., GTAs engaged in a higher percentage of open dialogue) [23]. Additionally, Becker *et al.* found biology GTAs used skills from pedagogical skill "drill" sessions when teaching

their students; however, GTAs' use of the pedagogical skills in the classroom was not stable [32]. Thus, further development and investigation of the impact of microteaching sessions on GTA and student outcomes is necessary.

3. Simulation is a promising tool for GTA preparation

The practice of training through simulation is embedded in situated cognition, which posits that learning and practice of knowledge are not separable, and practice of knowledge should occur in authentic environments [33]. Authentic simulations are leveraged for low-risk, low-cost training in multiple disciplines, including training medical students to perform surgical procedures [34–36] and training pilots [37]. Physics educators are likely familiar with PhET simulations, which are designed to support student learning of physics concepts [38,39].

While microteaching sessions have reported benefits for GTA training, they might not create authentic teaching experiences for STEM GTAs. He and Yan investigated the limitations of microteaching authenticity through interviews with preservice teachers participating in their English as a foreign language education program [40]. They found preservice teachers identified factors that decreased the authenticity of the microteaching experience, such as their peers' nonrealistic acting as high schoolers [40]. In their study about the role of technology in preservice teacher preparation, Brown similarly reported some preservice teachers find peers "playing" students is not as authentic of an experienced as they wished [41]. Brown suggests that technology can provide an alternative method to simulate more authentic classroom experiences [41].

The use of simulation can support new teacher preparation by providing beginning teachers with opportunities to practice teaching before entering the classroom, similar to microteaching. By interacting with a simulated class, teachers do not need to worry about the impact on students while trying a new teaching skill. In our study, we use TLE TeachLivE™ (TeachLivE) [42,43] to create a safe, authentic learning environment for practicing pedagogical skills. TeachLivE blends elements of virtual and physical reality to create a "mixed-reality" learning environment [42,43]; participants are in a physical room and can use "real" classroom artifacts (i.e., whiteboard) while interacting with avatar students in a simulated virtual classroom projected onto a screen as shown in Fig. 1 (described again in a later section).

Studies have reported the advantages of using the simulator for teacher preparation. In one study, a preservice teacher who used the simulator to practice teaching a lesson about technology to elementary students was observed to improve their interactions with real students during class discussions during the same lesson [43]. In an exploratory study, Chini *et al.* found physics learning assistants (LA) improved at least one skill noted in their self-reflections after one session with the simulator [44]. They also found participants reported the simulator created a realistic classroom environment including interactions with avatar students feeling natural over time and avatar-student personalities mirroring the personalities of real students [44]. In our previous studies about investigating mixed-reality simulation for possible use in GTA preparation, we found implementing simulator sessions into GTA training had a positive impact on chemistry and physics GTAs' use of cold or warm call in their classrooms (e.g., significantly larger use of cold or warm call for GTAs who participated in simulator training than nonsimulator training GTAs) [45,46]. Building on our previous findings, we continue to explore the use of the simulator as a tool for GTAs to practice asking questions to probe for student reasoning and integration of concepts or skills.

4. Questioning as a skill for improvement

The use of questioning supports student-centered instruction by shifting the focus of who supplies knowledge in the classroom from instructor to students. Questioning techniques can enhance student learning. For example, open-ended questioning can increase the diversity of

FIG. 1. Picture of a researcher (T. W.) interacting with the simulator. Left image: Virtual classroom projected on a screen in front of simulator user; Right Image: Researcher (T. W.) interacting with simulator in a physical room.

student ideas shared in the classroom [47], elaborative interrogation techniques have been correlated with increased test scores [48], and facilitators' use of Socratic questioning instead of confirmation when checking in with students has been linked to increased correct student responses to prompts about tutorial concepts [8]. Additionally, students who self-reported higher learning gains in a lab setting also perceived their GTA to be interactive and to ask thoughtful questions [49].

However, there is limited research on the impact of practicing questioning techniques during training on GTAs' implementation of questioning in their actual classroom. In our previous study, Wan *et al.* found a positive impact of implementing simulator sessions into GTA training on physics GTAs' use of questioning in general (e.g., GTAs in the simulator training program spent more time asking questions in front of the whole class than physics GTAs who did not receive same training) but did not center analysis on the use of a specific questioning technique [45]. To expand on our previous findings, we focus this study on investigating the impact of our training program on a specific questioning technique related to higher-order thinking named Stretch-It (described in Sec. III).

B. Factors to consider for GTA training design, implementation, and evaluation

We used Reeves *et al.*'s conceptual framework for GTA professional development evaluation and research (Fig. 2) to guide our study [16]. The framework suggests training design (e.g., training activities) and implementation (e.g., GTA buy-in) in parallel with GTA characteristics (e.g., previous teaching experience) impact GTA cognition (e.g., beliefs about teaching) [16]; GTA cognition paired with GTA characteristics influence GTA teaching practices (e.g., student-centered instruction). Student outcomes (e.g., student learning) are impacted by GTA beliefs and teaching behaviors [16]. Institutional factors such as departmental norms about teaching often influence the design of GTA training. For example, STEM disciplines might create a culture that views teaching as secondary to research [3]. However, professional development focused on teaching can help STEM GTAs create positive perceptions of teaching as a viable career (i.e., [31]) and useful to the development of other skills, like formulating richer research questions [50].

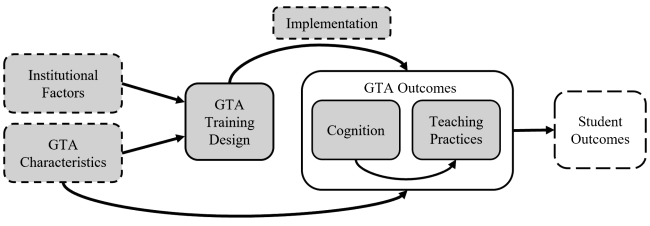

FIG. 2. Adapted version of the conceptual framework proposed in Ref. [16].

Guided by this framework, we designed a training program to encourage GTAs' use of pedagogical skills aligned with the instructional style of the target courses. We recognize GTAs have a variety of previous teaching experiences and beliefs about teaching, therefore, we have implemented a deliberate practice model with personalized feedback. In addition, we used a mixed-reality simulator to create an immersive, low-risk training environment. While in our larger project, we explore multiple relationships proposed by this framework, in this study, we investigate GTAs' participation (engagement) during the training and the impact of training on GTA teaching practices.

III. METHODOLOGY

A. Developing GTA training through action research

In line with action research [51,52], our research team consisted of the course designers and former instructors of the target courses. Based on these experiences, we recognized that the curricula required physics GTAs to use skills that were not supported by existing GTA training. We followed an iterative cycle with design, implementation, and reflection informed by GTA feedback of each training module as described in Geraets *et al.* [53].

B. Positionality

We recognize the variation of academic positions within our research group that contribute to different power dynamics between ourselves and our participants. C. M. D. executed several roles in the project and target courses: graduate student peer, training facilitator, and classroom observer. In addition, C. M. D. and T. W. led weekly preparation meetings for target courses. J. J. C. taught the pedagogy course, which most of the participating GTAs had either previously or were currently taking. A. A. G., C. A. N., and E. K. H. S. occupied similar roles in the chemistry department but were lesser known by and had less power in relation to the physics GTAs. Our presence as observers may have impacted GTAs' teaching practices, and the impact may vary across GTAs and GTA-observer combinations.

C. Target courses: Recitation and lab sections

The target courses are the two-semester introductory algebra-based physics "ministudio" (combined recitation and lab) sections [24,54]; ministudio sections are paired with a lecture component, and the grade earned in ministudio is incorporated into the overall course grade. The algebra-based physics sequence is primarily intended for life science majors. During each ministudio session, students are expected to work in small groups (3–4 students) on tutorial-style worksheets adapted from the "Tutorials on Physics Sense-making" [55], a group quiz, and an inquiry-style lab report based on the Investigative Science Learning Environment (ISLE) curriculum [56].

D. Participants

1. Presimulator training semester

Before implementing the simulator training intervention, the research team observed 8 (out of 13) ministudio GTAs teaching in their real classrooms during the Fall 2018. We refer to this sample as the baseline semester. Four of the GTAs were currently teaching Physics I, and the other four GTAs were teaching Physics II; in each group, one GTA had taught ministudio in a prior semester, while the other GTAs did not have prior GTA teaching experience. The GTAs varied in their teaching experience, gender, and nationality; however, demographic data were not collected, which is a limitation of this study since GTA characteristics may influence the impact of professional development [16]. All eight GTAs who consented to the study in the baseline semester consented to both researcher-created observation logs and audio recordings of their observations to be used for research purposes. Details of classroom observation data collection are described in a later section.

2. High-intensity simulator training semester

In Fall 2019, all ministudio GTAs participated in four training sessions with the simulator. We refer to this sample as the high-intensity simulator training semester [57]. There was no overlap between the presimulator training (Fall 2018) and high-intensity simulator training (Fall 2019) GTA cohorts. Of the16 GTAs, 13 (81% participation) consented to their classroom observation data (researcher-created observation logs and/or observation audio-recordings; consented activities are reported in Table I) and 11 GTAs (69% participation) consented to both simulator (training session video recordings) and classroom observation data to be used for research purposes. High-intensity simulator training participants varied in their experience teaching ministudio (displayed in Table I), as well as other teaching experience, gender, and nationality.

TABLE I. High Intensity simulator training semester participants

GTA	Ministudio course	GTA teaching experience
1	Physics I	New
2	Physics I	New
3[a]	Physics I	New
4[b]	Physics I	New
5[b]	Physics I	New
6	Physics I	New, other courses
7	Physics I	New, other courses
8[c]	Physics I	Ministudio
9[b]	Physics I	Ministudio
10	Physics II	New
11[a]	Physics II	New
12	Physics II	Mini-Studio
13	Physics II	Ministudio; other courses
14	Physics II	Ministudio; other courses

[a]Absent for one teaching observation.
[b]Classroom observation data only. GTA 4, 6, and 9 only had a log for classroom observations.
[c]Simulator session data only.

E. Preexisting GTA training

Graduate students new to the teaching assistant program attended a one-semester pedagogy seminar during their first semester as a GTA, led by J. J. C. Each week GTAs read articles about general education topics (e.g., Bloom's taxonomy, student mindset about learning, etc.), wrote reflections connecting their teaching experiences to the articles and participated in face-to-face Socratic discussions with other new GTAs.

In addition, all GTAs are required to attend weekly preparation meetings facilitated by a head GTA or training facilitator. Weekly preparation meetings for Physics I and II ministudio GTAs were facilitated by T. W. and C. M. D. during Fall 2018 and C. M. D. during Fall 2019. Meetings were run the same way in both semesters. During weekly preparation meetings, GTAs were encouraged to work together to discuss the tutorial worksheet and lab for the following week and to share their experiences in the classroom including common student ideas or mistakes.

Both the baseline and high-intensity simulator training cohorts attended a pedagogy seminar and weekly preparation meetings. The only difference between training programs is the addition of simulator training sessions in the Fall 2019, as described below. Thus, we make the assumption that impacts from other aspects of the GTA preparation program would be relatively similar, and that differences in GTA use of Stretch-It questioning are likely due to the simulator training sessions and/or individual GTA differences.

F. Simulator training: Teaching intervention

In the high-intensity simulator training model, the GTAs rehearsed in the simulator 4 times during a single semester. The first three simulator training sessions (STS1–STS3) focused on specific skills: (STS1) cold or warm calling (calling on nonvolunteering students to share ideas or answer questions [6]) and error framing (verbal statement framing student error has beneficial or natural to the learning process [6]); (STS2) questioning skills; and (STS3) group facilitation. In the fourth session (STS4), the GTAs were tasked with integrating the skills from the first three sessions into their interaction with the simulator. In this paper, we explore GTA engagement with the training by tracking their use of the target questioning skill (Stretch-It) during (1) the session solely focused on GTAs' rehearsal of questioning skills (STS2), and (2) other simulator sessions (STS1, STS3, and STS4).

To create a realistic experience with the simulator, we included common student ideas and reasoning in avatar-student dialogue, and we leveraged the personalities of the

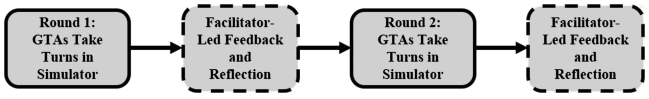

FIG. 3. Schematic of the cyclical procedure for a simulator session.

five avatar students (Kevin, CJ, Ed, Maria, and Sean) to create dysfunction within student groups (for STS3, focused on group facilitation). Additionally, we framed each simulator session to be a snapshot of teaching in either recitation or lab such that avatar students worked in groups on activities similar to the ministudio curricula. In the questioning skill session (STS2), we implemented an activity from the "Thinking Critically" lab curriculum [58,59]. While there is a large difference in class size between the simulator (five avatar students) and the real classroom (up to 32 students), we do not expect the class size to impact GTAs' rehearsal of Stretch-It questions in the simulator or use of Stretch-It in the classroom. It is possible features of the simulator will support GTAs practice of the skill regardless of this limitation. For example, in our pilot sessions with experienced GTAs, GTAs most frequently commented that it was easier to use students' names (as in cold calling) with the smaller size and seating chart provided in the simulator. GTAs were given about 7 min to perform check-ins with two student groups, which more closely models the real classroom teacher-student ratio.

Before each session, we provided GTAs with descriptions and examples of the target pedagogical skills and a brief description of the simulator and avatar students. Simulator sessions followed a cyclical procedure of deliberate practice: approximately 7-min sessions to practice teaching in the simulator (round 1 and 2) with facilitator-led feedback and reflection in between rounds, as depicted in Fig. 3. During feedback and reflection, facilitators (coauthors of this study) led discussions about the GTAs' use of the skill in their first round and provided feedback. Facilitator-supplied feedback was semistructured. Facilitators began with common prompts created to direct GTA attention to their use of the target skill during the session (i.e., "Do you feel like you were able to use the target skill?") and their use of the skill in their actual class (i.e., "How would you use this in your class?"). Then, facilitators followed up on what occurred during each session and GTA responses. GTAs typically participated in groups of three such that each GTA individually rehearsed in the simulator while the other GTAs in their group watched them interact with the simulator. We have provided one round of dialogue for one GTA in Appendix A as an example of how GTAs interacted with the avatar students.

G. Operationalizing questioning skills

We operationalized the term questioning skill to be posing Stretch-It questions to (avatar-) students. We selected Stretch-It as a focal skill because it provides a way for GTAs to build on student ideas, which is a goal of the instructional materials in the target courses. Prior GTAs had expressed difficulty responding to students without directly (dis)confirming the student's idea or explaining the content. Stretch-It questions are operationalized by two

TABLE II. Questioning skills presented in simulator intervention. Example quotes were taken from GTA simulator training sessions and modified for clarity and brevity.

	Categories	Sub-categories
Stretch-It question types	Explain Logic: the GTA asks a student to explain the reasoning behind an answer or idea the student shared	Ask for Work: The GTA asks a student to explain how they got their answer *Example: "How did you measure the period?"* Ask for Evidence: The GTA asks a student why their answer is valid (or not) *Example: "How do you determine if your data is more reliable than the other group's?"*
	Follow-Up: the GTA asks a student to stretch the boundaries of their knowledge and checks for integration	Ask About an Analogous Situation: The GTA asks a student to relate the concept or skill to a different context *Example: "Think about a dart board. If you have a really accurate collection of darts, they've all like gotten really close to the center. Right? So, what do we call it when all the darts have hit the same spot, but it's not the center?"* Ask for another way to answer: the GTA asks a student to explain or answer in a different way. *Example: "That sounds very textbook. Could you say that without so much technical jargon?"*

main categories, each with two subcategories. "Explain Logic" questions prompt students to provide reasoning for their shared idea by the GTA either "asking for work" or "asking for evidence" [6]. "Follow-Up" questions encourage students to apply their original idea or understanding in new ways. A GTA can ask a Follow-Up question by asking students about an analogous situation or using a different method to answer the original question [6]. Definitions and examples of Stretch-It question types are provided in Table II.

Besides Stretch-It, instructors use a variety of questions that vary in the cognitive activity required to respond. For example, instructors may ask rhetorical questions, which do not require a response at all, or instructors may ask "leading" questions, which scaffold the cognitive work while providing an opportunity for students to respond. Exploring all question types is beyond the scope of this project, so we limit our analysis to Stretch-It questions, which fit both GTAs' needs and the curricular goals. We note that this analysis does not capture all of the valuable questions GTAs asked and does not assess the quality of questions beyond meeting the definition for Stretch-It.

IV. DATA COLLECTION AND ANALYSIS

A. Simulator sessions with the mixed-reality classroom simulator

In the high-intensity simulator training semester, we analyzed simulator sessions for 11 physics GTAs. About 41 simulator sessions were audio and video recorded amounting to a total of 586 min. All simulator sessions were transcribed by C. M. D. Co-coders C. M. D., A. A. G., and E. K. H. S. participated in a training session before coding the simulator training data. Here, a small portion of the data (∼14 min in total; portions of four different GTAs' sessions) was used for co-coders to practice implementing the *a priori* codes (described in Table II; codebook) by independently finding examples of Stretch-It questions and discussing examples until agreement was reached [60]. After the practice round, A. A. G. and E. K. H. S. coded two different sets of simulator data for instances of Stretch-It, totaling 41 and 16 min, respectively, and C. M. D. independently coded both sets of data (∼10% of the simulator data). The grain size for coding simulator training data was GTA turn-of-speech (when a GTA was talking with/at avatar students). We explored interrater reliability (IRR) by calculating Gwet's AC1 values for pairs of coders; Gwet's AC1 is a metric that can be used to interpret IRR and is robust for data with low-trait prevalence [61,62]. Averaged Gwet's AC1 values for all four subcategories of Stretch-It ranged between 0.93 and 0.99 before the discussion and 0.97 to 1 after the discussion [62]. Afterward, C. M. D. coded the remaining simulator training data (∼529 min) using the same coding method.

B. Classroom observations

In the baseline semester, we observed eight ministudio GTAs teaching in their classroom during five lessons (a total of 32 observations). In the high-intensity simulator training semester, we observed 13 ministudio GTAs teaching in their classroom during 3 lessons (a total of 37 observations). Both intervention and observation timelines are shown in Fig. 4.

Classroom observations were conducted by C. M. D., A. A. G., T. W., and C. N. N. using a modified version of the Laboratory Observation Protocol for Undergraduate STEM (LOPUS) to capture GTA and student behaviors in the classroom [63] as described in Wan *et al.* [45]. Here, we focus specifically on the codes for Stretch-It. During observations, Stretch-It was coded at the category level (Explain Logic and Follow-Up) to reduce cognitive load on observers.

Like simulator training data, we calculated averaged Gwet's AC1 values across pairs of coders for Explain Logic, 0.94 before discussion and 1 after discussion, and for Follow-Up, 0.99 before discussion and 0.99 after discussion. We also calculated Gwet's AC1 values for two GTA behaviors mentioned later in this section, Posing Question (PQ) and One-on-One TA Posing Question (1o1_TPQ). Averaged Gwet's AC1 values for PQ and 1o1_TPQ were 0.74 and 0.95, respectively. Values between 0.61 and 0.8 can be considered moderate agreement while values above 0.81 can be interpreted as near-perfect agreement [62].

To assess the impact of the simulator training on GTAs' implementation of Stretch-It questions in the classroom, we compared high-intensity simulator training classroom observation data to baseline classroom observation data for each category of Stretch-It [64]. However, we did not originally code for Stretch-It in the baseline semester observations, so we revisited that data for this analysis. C. M. D. used the observation logs to identify 2-min intervals where the GTA asked a question, as indicated by the PQ or 1o1_TPQ codes. Then, C. M. D. listened to the corresponding 2-min interval in the audio file and coded for Explain Logic and Follow-Up Questions. C. M. D. conducted this analysis after demonstrating acceptable

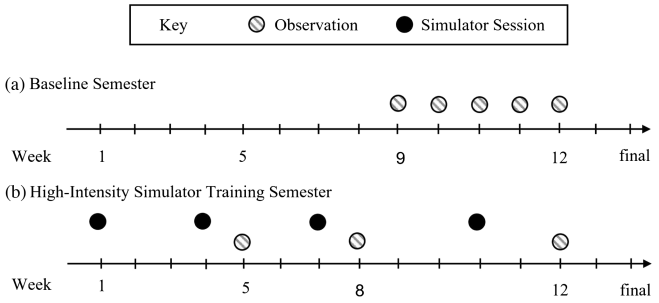

FIG. 4. Timelines for each cohort. (a) Top: Timeline of baseline semester (Fall 2018) and (b) Timeline of high-intensity simulator training semester (Fall 2019) intervention.

IRR with the research team on the high-intensity training semester classroom observations and simulator sessions.

V. FINDINGS

A. GTAs' use of stretch-it during high-intensity simulator training

Figure 5 displays GTAs' use of all four Stretch-It categories during STS2 for both rounds GTAs interacted with the simulator. On average, GTAs had about 25 turns of speech per 7-min round. The total turns-of-speech GTAs asked a Stretch-It question increased from 22 to 33 turns-of-speech between round 1 and round 2. Six GTAs increased the number of turns-of-speech in which they asked Stretch-It questions from round 1 to round 2, while two GTAs asked Stretch-It questions in the same number of turns-of-speech in both rounds and three GTAs decreased the number of turns-of-speech in which they asked a Stretch-It question across rounds. We observed 9 of 11 GTAs use Stretch-It in round 1, and all 11 GTAs use Stretch-It in round 2. Thus, we found that after receiving feedback, more GTAs used Stretch-It questions. Additionally, both the quantity of Stretch-It questions and the variety of Stretch-It question types increased. We infer that rehearsal in the simulator with the opportunity to observe peers and receive personalized feedback supported GTAs to implement Stretch-It questions.

1. GTAs used different types of stretch-it questions in different rounds

In round 1, we found "Ask for Work" to be the most common type of Stretch-It question posed across Physics I and II GTAs (7 of 11 GTAs) followed by "Ask for Evidence," which was used by mostly Physics II GTAs (four GTAs), and "Ask for Another Way to Answer," which was used by only Physics I GTAs (two GTAs). We did not observe any GTA using "Ask about an Analogous Situation" in round 1.

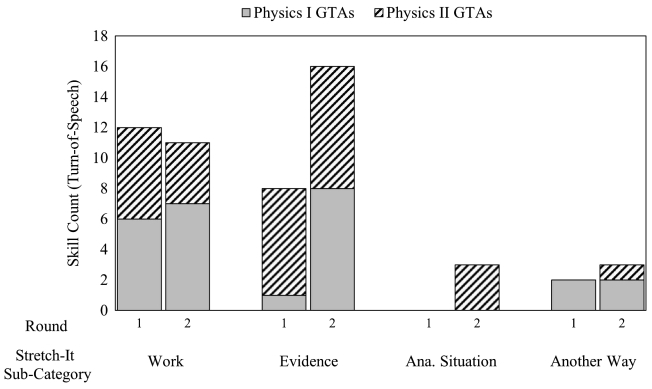

FIG. 5. Stacked, clustered bar charts document GTAs' use of "Stretch-It" during the questioning skills session (STS2). The horizontal axis represents the four subcategories of Stretch-It, and the vertical axis represents the count of (turns-of-speech) for each Stretch-It subcategory. Physics I and II GTAs' use of each subcategory of Stretch-It is distinguished in the chart according to box fill. There are two columns for each subcategory indicating the count of Stretch-It sub-category for round 1 (left column) and round 2 (right column).

After facilitator-led feedback and reflection, eight GTAs used both Ask for Evidence and Ask for Work questions. Physics I and II GTAs' use of Ask for Work remained consistent across rounds. Ask for Evidence was used more frequently than Ask for Work, a reversal compared to round 1. We observed Physics I GTAs increase their use of Ask for Evidence from round 1 to round 2 while Physics II GTAs kept their use of this subcategory consistent across both rounds. We also observed three Physics II GTAs use Ask about an Analogous Situation during round 2 (no GTAs used this question type in round 1). Thus, we infer feedback and the opportunity to observe peers promoted GTAs to ask different types of questions.

2. GTAs continue to use stretch-it in subsequent simulator sessions

We compared GTAs' performance before and after the simulator session focused on Stretch-It. The count (turn-of-speech) for each subcategory of Stretch-It for each of the four simulator sessions is displayed in Fig. 6. We combined counts from round 1 and round 2 and present a total count of turns of speech with a Stretch-It question per simulator session.

GTAs asked the most Stretch-It questions (total of 55 turns of speech) in the session focused on asking Stretch-It questions (STS2) and the least (total of 13 turns of speech) in the session focused on cold call and error framing (STS1). This finding is expected since STS2 focused specifically on questioning skills, and STS1 occurred before the GTAs practiced questioning skills in the simulator.

We also observed that Physics I and II GTAs continued to ask Stretch-It questions during STS3 (Group Management;

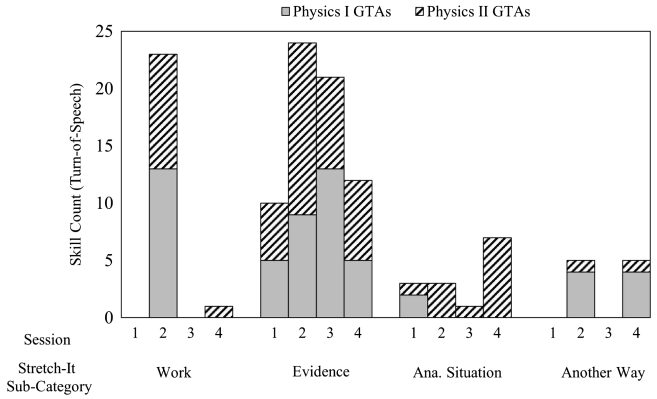

FIG. 6. Stacked, clustered bar charts document GTAs' use of Stretch-It across all four simulator sessions. About 10 GTAs participated in session 1, 11 GTAs participated in sessions 2 and 3, and 9 GTAs participated in session 4. The average simulator session was 25 turn of speech.

a total of 22 turns of speech) and STS4 (Integrated Skills; a total of 25 turns of speech) when the focal skill was no longer Stretch-It. While the number of turns of speech decreased from STS2 (Stretch-It), they were close to twice the total turns of speech in STS1. Across all sessions, GTAs most frequently used Ask for Evidence questions. We also observed at least one GTA using Ask About Analogous Situation questions during each simulator session. These results show that GTAs continued to use Stretch-It when they were not explicitly prompted to practice Stretch-It, especially Ask for Evidence.

3. GTAs use all subcategories of stretch-it during sessions with feedback about that skill

During STS2 and STS4, GTAs received tailored feedback about their use of Stretch-It in the simulator. We found GTAs either increased their use of Stretch-It or used different categories of Stretch-It during round 2 compared to round 1 for STS2. Furthermore, we observed GTAs to continue to use a variety of Stretch-It categories in STS4, unlike STS1 or STS3 where feedback was not specific to the use of Stretch-It. Additionally, we observed GTAs to only use Ask for Work and "Ask for Another Way" during STS2 and STS4; GTAs' use of Ask for Work was substantially higher in STS2 than in STS4. This finding implies that facilitator-led feedback and reflection were important to the GTAs' use of Stretch-It in the simulator. For example, even though GTAs posed a Stretch-It question for a similar number of turns of speech during STS3 and STS4, GTAs asked all four subcategories of Stretch-It (at least one GTA per subcategory) during STS4 in contrast to STS3, where GTAs predominantly used Ask for Evidence questions. It is also possible GTAs practiced different types of Stretch-It during STS4 (like STS2) because they were prompted to practice target skills they wished to revisit.

B. Investigating GTAs' use of stretch-it: Explain logic in the classroom

Since we broke down Stretch-It into two main categories when coding in the classroom, we present and discuss GTAs' use of Stretch-It: Explain Logic and Stretch-It: Follow-Up separately.

Here, we present the use of Stretch-It in the classroom for the high-intensity simulator training GTAs as well as the baseline GTAs. On average, the length of a classroom observation during the baseline semester and high-intensity simulator training semester was 155 min (57.4 two-minute intervals) and 151 min (75.5 two-minute intervals), respectively.

1. Describing GTAs' use of explain logic in the classroom

To visualize GTAs' use of Explain Logic during the baseline semester and high-intensity simulator training semester, we created violin plots for four different collections of classroom observation data (baseline and three high-intensity training observations), displayed in Fig. 7. Dot plots were included to show the distribution of individual observations. Since we did not have an intervention in the baseline semester, we grouped all observations for that semester. Mean and median values for each collection of observations are displayed in a box above each violin plot in Fig. 7.

Violin plots, named after the musical instrument, are similar to boxplots and provide a more in-depth representation of the spread of data by using kernel density

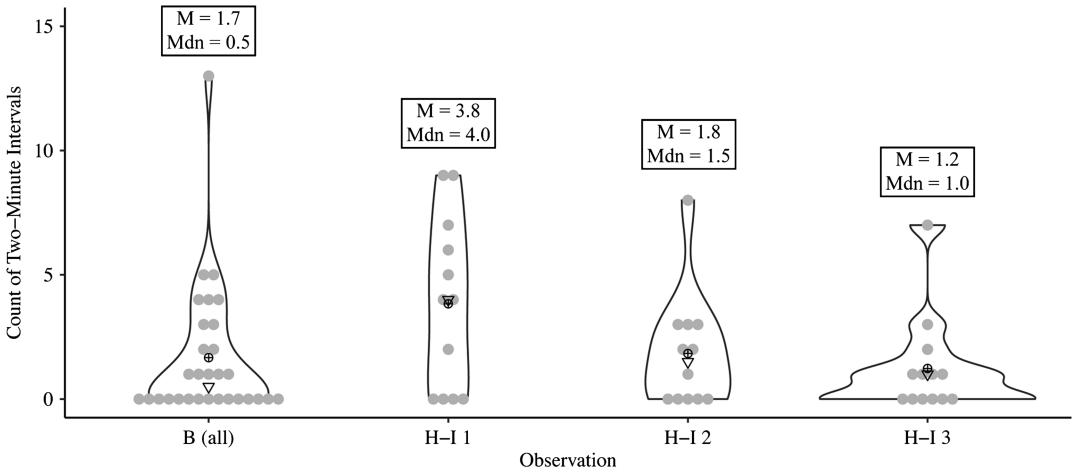

FIG. 7. Violin plots visually represent the distribution of GTAs' use of Explain Logic in the classroom. Each violin plot represents the density of each collection of observations. Dot plots overlayed on each violin plot represent the distribution of individual GTA observations. Note that the baseline group contains multiple data points from the same set of eight GTAs, whereas each high-intensity training semester observation contains one data point from each GTA observed teaching that lesson. B (all)—all baseline observations, H-I 1—high-intensity simulator training semester observation 1, H-I 2—high-intensity simulator training semester observation 2, and H-I 3—high-intensity simulator training semester observation 3.

estimation to create a density trace or curved shape (instead of a box) [65]. Since our sample sizes are small which might inflate or deflate any statistical findings, we do not attempt to claim statistical significance with our findings. Instead, we use the violin plots and median and mean values to make sense of the qualitative differences we observed. We generated violin plots along with each corresponding dot plot using the ggplot, geom_violin, and geom_dotplot functions in the R package "ggplot2" [66,67].

In the baseline semester, we observed no Explain Logic questions in half of the observations, as indicated by the dots on the horizontal axis and the width of the bottom of the violin plot. In baseline observations where Explain Logic questions were observed, GTAs tended to use Explain Logic in one to five two-minute intervals. Due to a hurricane at the beginning of the Fall 2019 semester, we do not have an observation for the high-intensity simulator training GTAs before STS2 (Stretch-It). Therefore, we compare baseline semester GTAs' use of Explain Logic to high-intensity simulator training semester GTAs' use during each observation.

We find that high-intensity simulator training GTAs' use of Explain Logic in the observation immediately following STS2 was markedly different than the baseline GTAs' use of Explain Logic as well as the use in subsequent observations by the high-intensity simulator training GTAs. The shape of the violin plot for high-intensity observation 1 (H-I 1) nearly resembles a uniform distribution ranging from zero to nine two-minute intervals, while the baseline observations (B(all)) and high-intensity observation 2 (H-I 2) and 3 (H-I 3) shapes resemble a skewed distribution centered around "0." When comparing median and mean values, H-I 1 (median = 3.83 two-minute intervals and mean = 4.00 two-minute intervals) has larger values than subsequent observations in the same semester. This finding suggests the simulator session focused on using questioning skills (STS2) had an immediate impact on GTA questioning behavior in the classroom as more GTAs asked more Explain Logic questions in this observation (H-I 1).

However, we find GTAs collectively decreased their use of Explain Logic in later observations during the high-intensity training semester given median values decreased. Additionally, the shapes of the violin plots for H-I 2 and H-I 3 progressively start to resemble the same shape as the violin plot for the baseline semester. While H-I 1 took place the week following GTA participating in the simulator session focused on using questioning techniques, these two observations took place one to two weeks after GTAs participated in simulator sessions focused on group management and using all target pedagogical skills. Therefore, we observed an immediate impact of the simulator training on GTAs' use of questioning in the classroom, but the implementation "boost" to the skill was not stable.

C. No measured impact of training on GTAs' use of follow-up

Similar to our analysis of GTAs' use of Explain Logic, we created violin plots overlayed with dot plots for the classroom observation data from the baseline and three high-intensity simulator training observations, displayed in Fig. 8. However, unlike GTAs' use of Explain Logic in the classroom, the simulator training does not have a clear impact on GTAs' use of Follow-Up. Median values for all observations for both cohorts were zero, and all data points

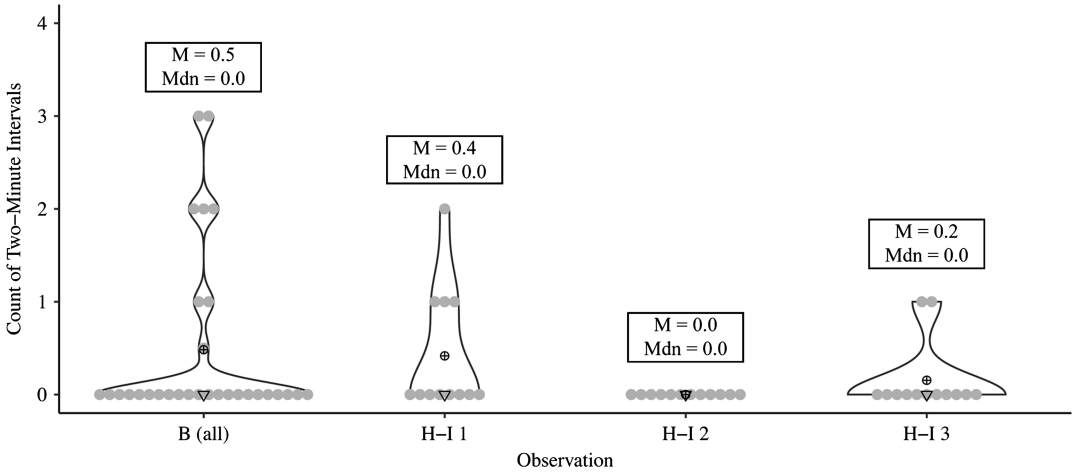

FIG. 8. Violin plots visually represent the distribution of GTAs' use of Follow-Up in the classroom. Each violin plot represents the density of each collection of observations. Dot plots overlayed on each violin plot represent the distribution of individual GTA observations. Note that the baseline group contains multiple data points from the same set of eight GTAs, whereas each high-intensity training semester observation contains one data point from each GTA observed teaching that lesson. B (all)—all baseline observations, H-I 1—high-intensity simulator Training semester observation 1, H-I 2—high-intensity simulator training semester observation 2, and H-I 3—high-intensity simulator training semester observation 3.

above zero are outliers (i.e., more than 1.5 times beyond the interquartile range. It is also worth noting that a higher percentage of GTAs from the presimulator training cohort used Follow-Up in the classroom than the high-intensity training cohort (63% compared to 46%). These findings imply that GTAs, regardless of training, infrequently used Follow-Up in their classroom.

VI. DISCUSSION AND IMPLICATIONS

A. GTAs benefit from facilitator-led feedback and reflection

We found GTAs used all categories of Stretch-It when we guided GTAs' reflection and provided specific feedback about their use of Stretch-It in the simulator. This finding implies that guided feedback and reflection might be crucial to GTAs' implementation of evidence-based practices during training sessions and in the classroom. However, we also observed GTAs decreased their use of Explain Logic in the classroom over time, similar to the finding by Becker *et al.* [32], and GTAs' use of Follow-Up did not fit a trend. Therefore, we suggest that GTA training developers and facilitators create opportunities for GTAs to receive continuous feedback, such as practice sessions using pedagogical skills outside of the classroom, and lead discussions with GTAs about their use of target skills in the classroom.

Based on our findings, we plan to incorporate more direct feedback about the GTAs' use of Stretch-It in each simulator session and include feedback from GTAs to improve the training program for future iterations. We also plan to create a more holistic approach to preparing GTAs by overlapping all facets of training received. For example, we will include commentary on their use of the skill during classroom observations in feedback during simulator sessions, discuss successes and challenges with using Stretch-It in the classroom during weekly preparation meetings, and prompt GTAs to read and write reflections about Stretch-It as part of their reading assignments in the pedagogy course.

B. Follow-up questions might be difficult to ask

We did not find a direct impact of the simulator training on GTAs' use of Follow-Up in the classroom. Overall, Explain Logic was used more frequently than Follow-Up during both simulator sessions and classroom observations. It is possible that GTAs found Explain Logic questions less difficult to ask than Follow-Up questions, as suggested by GTA 10 during STS2.

> GTA 10: "Um so some of them [questions] seem like things I do more naturally 'cause it will actually be about something they did or something, but things like asking it, to integrate it, asking them to integrate it to a related skill or like the new setting like those require more thought… Some of the earlier things on this list [referring to handout about Stretch-It] are easier to come up with but like I want to try to fit in something like with like integrated skill or like applied setting."

While we provided GTAs with examples of how to ask Follow-Up questions in the simulator, it is possible GTAs needed more examples, scaffolding, and time to become proficient in using these types of questions in the simulator and their real classroom. Since asking about analogous situations and for other ways to explain supports higher-order thinking, we suggest providing GTAs with more examples of Follow-Up in addition to asking GTAs to think of analogies for common themes in their physics course.

C. GTAs might be using stretch-it as a group management strategy

We observed GTAs continue to use Ask for Evidence during STS3 (Group Dynamics). During this session, facilitators instructed GTAs to practice supporting healthy group interactions while interacting with student groups with various dysfunctions (e.g., one avatar student dominated the discussion).

Some GTAs used Explain Logic questions to hear from each group member. For example, after hearing from Sean and Kevin that they did not work with Maria, GTA 6 attempted to encourage communication between the avatar students by asking Maria for the reasoning behind her idea (dialogue included in Appendix B). A similar strategy can be seen in a video published through Periscope, a repository of free instructor training materials. In the Periscope lesson, "How can I bring out students' idea" paired with "Episode 101: Depth," the instructor "Levi" listens to everyone's idea in the group before insinuating the group is in disagreement [68]. It is possible GTAs were using a similar strategy and using "Ask for Reasoning" to hear from everyone in each avatar-student group. Future work will investigate the use of questioning when GTAs facilitate group work.

D. Some GTAs voiced concerns about student frustration when using stretch-it in the classroom

Our results suggest that providing GTAs with opportunities to rehearse asking questions in a low-risk environment had an immediate positive impact on GTA use of Explain Logic in the classroom, as supported by the model of situated cognition. However, GTAs did not consistently use Stretch-It in the classroom throughout the high-intensity simulator training semester. Other factors like GTA teaching experiences also impact GTA cognition and teaching practices [16]. Here, we discuss two GTAs' concerns about student resistance when using Stretch-It in their actual classroom. Their concerns might also

be a reason why we did not find the impact we expected (e.g., GTAs' continued use of Stretch-It throughout the semester).

During reflection and feedback sessions, we asked GTAs to think about their performance of the skill in the simulator and how they might apply it (or are applying it) to their interactions with real students. Two GTAs (GTA 13 and GTA 1) voiced their concerns in detail about using Stretch-It in the classroom. Both GTAs used Stretch-It during STS1 and participated in different simulator sessions throughout the semester.

After their final simulator round for STS2, GTA 13 responded to the facilitator's prompt about using Stretch-It in their real classroom: "*Yeah actually I noticed something yesterday, too. We had this worksheet that compared the fans, the airfield to the electric field, and it really helped them. So, this is kind of similar that you can apply to more familiar situations and compare it with, it really helps. It's something I will try more*." Thus, GTA 13 found Stretch-It questions, in particular Follow-Up, to be useful for student learning since they observed students were supported by similar question types in the tutorial worksheet. In addition, we observed GTA 13 use both types of Stretch-It during observation 1 (H-I 1). However, we did not observe GTA 13 use Stretch-It in the classroom for the remainder of the semester. During STS4, we asked GTAs about their goal for the session, and GTA 13 stated, "*Yeah. I will probably focus on the Stretch-It things because I don't know, sometimes it annoys them, but I know it's very, very important*." Here, GTA 13 still believed Stretch-It was useful for student learning, but their students' response as annoyance may have led them to stop using Stretch-It in the classroom.

GTA 1 shared a similar experience as GTA 13 with student frustration associated with asking Stretch-It questions in class. After round 2 during STS2, we asked the GTAs if they would consider using Stretch-It in the classroom. GTA 1 responded, "*(laughs) I tried it now [Stretch-It] and they get frustrated with me*." When asked to elaborate on what they mean, GTA 1 added, "*I'll try to get them to see what they are doing wrong but then they are just like 'I don't know what I am doing so I can't tell you what I am doing wrong*'." While GTA 13 stopped using Stretch-It in the classroom, GTA 1's use of Explain Logic did not fit any trend. In fact, we observed GTA 1 use Explain Logic more frequently during the last classroom observation
(H-I 3) than the first classroom observation, regardless of experiencing continued student frustration, as mentioned before their first round in STS4, "*I usually do a pretty good job with stretching it. I just try, even if they get it right, I ask them 'why is that? Tell me'. My students get mad at that because [they are like] 'why are you doing this if I am right?' (small laugh)*." Despite their concern with student frustration about their use of Stretch-It in the classroom, GTA 1 continued to use Stretch-It.

Based on the cases of GTA 13 and GTA 1, we infer that although GTAs might experience similar situations while implementing Stretch-It in the classroom, each GTA might have a different reaction to student resistance. The curricular shift to support inquiry and sense-making might be unexpected for students who expected their physics course would be noninteractive [69]. The varying cases of GTA 13 and GTA 1 highlight a flaw in current GTA professional development with respect to how to react to student resistance. As suggested by Reeves *et al.*, GTA characteristics, including teaching experience, influence GTA beliefs about teaching and their teaching practices [16]. Thus, if a GTA has a negative experience with using a pedagogical skill in the classroom, they might associate their student's dissatisfaction with that skill, possibly causing GTAs to not use the skill again. This finding calls for further investigation into how and why GTAs may or may not implement pieces of professional development in their classroom. While the current study does not report on GTA personal characteristics like gender, race or ethnicity, or nationality, we suspect these and other dimensions of identity impact GTA-student interactions and GTA use of questioning.

Furthermore, we did not discuss ways GTAs can manage student resistance or frustration in the classroom, even though we were introducing GTAs to pedagogical skills that might violate how students expect their physics GTA to interact with them. We strongly suggest GTA training developers and facilitators incorporate discussions focused on strategies to manage student resistance with respect to evidence-based teaching practices and active learning instruction with GTAs. In future iterations of this program, we plan to include discussions of how to mitigate student frustration with pedagogical skills like Stretch-It.

VII. LIMITATIONS AND FUTURE WORK

There are a few possible reasons why we found these results. First, it is possible GTAs are implementing other targeted skills in the classroom given that the training focused on using and integrating multiple evidence-based practices in the GTAs' teaching repertoire. Also, GTAs might have interpreted Stretch-It in a different way than facilitator-provided material and therefore might have been using their own contextualized version of Stretch-It in the classroom (i.e., we found GTAs might contextualize target skills differently than facilitators [46]). Finally, it is possible GTAs are using Stretch-It in the classroom, but our current analysis does not capture every moment or pedagogical choice a GTA might make in the classroom. Future work might include collecting GTA artifacts like written reflections on GTA perspectives of the skill used in the simulator and classroom.

Simulator training sessions were designed to simulate a small portion of the ministudio class in order to focus GTAs' attention on the practice of target pedagogical skills within the context of the ministudio curriculum. To achieve this goal, designers limited the amount of time each GTA spent teaching in the simulator to 7 min per round (two rounds per session) in order to make time for GTAs to observe other GTAs' teaching and receive feedback before trying again. Future work will continue to investigate the optimal amount of time spent in the simulator needed to impact GTA teaching behaviors in their actual classroom. Additionally, GTAs interacted with fewer avatar students during simulator training sessions compared to students enrolled in their classes (∼32 students per class) as part of a limitation of the simulator. This limitation was considered during the design of the simulator sessions such that designers chose pedagogical strategies that complemented the simulator environment. However, it is possible the use of simulator training sessions alone is not sufficient to support some crucial aspects of GTA training like managing large amounts of students in the classroom. Instead, simulator training sessions can be paired with other training models like GTA preparation meetings and pedagogical seminars. Future iterations of this work will include the investigation of how different branches of congruent training support various GTA needs.

Additionally, we explored incorporating sessions with a mixed-reality classroom simulator into physics GTA training because of the promising role technology might play in preparing future teachers [41–44]. However, we recognize not every institution has access to such technology when training STEM GTAs. Future work will compare the use of incorporating simulator sessions with incorporating traditional microteaching sessions (as described in [26]) in STEM GTA training.

We chose to live code classroom observations instead of using video analysis to document GTA behaviors in real time; this allowed us to collect data for more GTAs as some GTAs were more comfortable with consenting just their observation log (not audio recorded) and video recording would require student consent. During classroom observations, GTAs wore a lapel microphone connected to an audio transmitter while researchers listened to GTA-student conversations through headphones attached to a receiver. An audio recorder was fed into the loop to record. We encountered issues, such as audio interference, wire connections coming loose in the middle of an observation, and difficulty hearing over noisy student groups, which might have led to an observer missing a code at a two-minute interval. This could lead to an underrepresentation of Stretch-It in our findings, as also discussed in Ref. [45].

Future research should explore the use of a simulator for training physics GTAs who teach in student-centered recitation and lab sections across institutions to allow investigation of contextual variables [16]. Future research should also disaggregate findings by GTA characteristics, which could affect both GTA receptiveness to the skills and student response to GTA use of the skills.

The ministudio sections are not stand-alone courses and were paired with lecture sections; ministudio grades counted an average of 20% of the lecture final grade. It is possible GTAs were influenced by the lecture instructor to implement different teaching practices or activities during either the recitation or lab portion as was demonstrated in a multitiered professional development project for calculus instructors and GTAs [70]. For example, we observed GTA 1 go over a practice test with students after they asked what students wanted to work on during recitation; the practice test was taken from the lecture course website.

We also recognize each lesson might provide GTAs with different opportunities to use Stretch-It in their ministudio section. However, we find it difficult to disentangle the culture and the lesson from our findings. Future work will aim to incorporate these two pieces into our investigation.

Since physics GTAs might have participated in simulator sessions with chemistry or mathematics GTAs, it is possible watching GTAs in other disciplines influenced physics GTAs' rehearsal of and discussion of the skill during training sessions or use of the skill in the classroom. We find it difficult to disentangle this aspect of the training from our findings. However, future work will investigate the benefits and barriers to training GTAs from multiple STEM disciplines together on their responsiveness to the training. This is especially important to investigate since it might be possible to create a more cohesive learning experience for STEM students if the GTAs in each of their STEM labs or discussion classes were trained to use pedagogical skills supportive of students learning.

Finally, it is possible international graduate students were hesitant to participate in the training for research purposes during the beginning stages of the project. In future work, we plan to explore possible reasons international GTAs might hesitate to participate in training programs and how to ensure international GTAs feel supported by training initiatives.

VIII. CONCLUSION

In summary, we observed GTAs increase or shift their use of Stretch-It across rounds during the simulator session focused on using Stretch-It. In addition, we observed all four subcategories of Stretch-It to be used by GTAs during simulator sessions with focused feedback and reflection about Stretch-It. We also observed most GTAs use "Ask for Evidence" during the group dynamics simulator session, suggesting GTAs might have used that subcategory of Stretch-It as a strategy for group management.

In addition, we found simulator training focused on the use of questioning to impact GTAs' use of questioning in the classroom. Specifically, we observed GTAs ask more Explain Logic questions during the classroom observation

immediately following the practice of Stretch-It in the simulator and when compared to baseline semester GTAs' use of Explain Logic, which suggests more GTAs are intentionally using Explain Logic in the classroom because of the training. However, the simulator training did not impact GTAs' use of Follow-Up in the classroom; GTAs might randomly use Follow-Up because of the curriculum or other factors such as GTA characteristics or cognition.

From our findings, GTAs might have a difficult time becoming proficient with questioning techniques, which may be an expected skill in student-centered curricula. While we found the simulator training influenced GTAs to ask more Explain Logic in their classroom, most GTAs did not maintain the usage of the skill in their classroom in later classroom observations. STEM departments should dedicate time to helping GTAs create Stretch It-type questions and practice asking them during GTA training if they have selected to use such a curriculum.

Due to the COVID-19 pandemic, our department changed the curriculum for the physics ministudios to require less face-to-face interaction for several semesters. This curricular change decreased the relevance of and opportunities to use the target teaching skills. Thus, we cannot presently speak of the long-term impact of the training on our department.

ACKNOWLEDGMENTS

We would like to thank all GTAs who participated in our study as well as the larger project. This study was funded by the NSF Grant No. 1725554.

APPENDIX A

The following is a snippet of dialogue between GTA 13 and the virtual class during their simulator session. The context of the session included avatar students working in groups to collect and analyze data to determine the period of a pendulum, a lab from the "Thinking Critically" curriculum. Avatar students CJ and Ed worked in a group while Kevin, Maria, and Sean worked together.

GTA 13: *Start classroom. [pause] Uh, so hi everybody. Let's start the class.*
Sean: *Hi.*
GTA 13: *Uh, how are you, how, what's going on?*
Sean: *Pretty good! How are you?*
GTA 13: *Good. Thank you. So, I know that you guys completed the experiment. So how about tell me, uh, you guys, how did you do the experiment? Sean, you want to start?*
Sean: *Uh, sure. Uh, where do you want me to begin with like technique or results or where do you wanna start?*
GTA 13: *The procedure. How did you precede the experiment?*
Sean: *Okay. So, um, what we did was we have, of course we have to measure at two angles at 10 degrees or 20 degrees. So, what we did was we measured 20 swings five times at 10 degrees, and then we measured it 20 swings, five times at 20 degrees.*
GTA 13: *Alright. So, how about the other group? CJ, How did you measure the time?*
CJ: *Um, well we measure 10 at 10 degrees and 10 at 20.*
GTA 13: *All right. So, we have two different, uh, periods. So, CJ and Ed, measured time, 14, 10 code 10 swings, and Kevin and Maria, Sean, they measured the time for 20 swings. Right? So.*
Kevin: *Yeah, we did a lot more than they did.*
GTA 13: *So, Oh yeah, yeah. If you really think so. Oh yeah, it's double, right? Uh, so, uh, uh, Kevin. So, are you thinking, do you think that your data are reliable?*
Kevin: *Um, well we've got a pretty high t score, so I'm thinking our data. There must, there must be something wrong with it.*
GTA 13: *Okay. So, what is the t-score?*
Kevin: *The t-score we got was 3.71.*
GTA 13: *Okay. So, CJ and Ed, do you think your data are reliable, more reliable than them?*
CJ: *Yeah.*
GTA 13: *Uh, how do you determine your data is more reliable?*

[The GTA continued to interact with avatar students until the timer went off.]
Facilitator: *Stop Classroom. … You just got that one in [Stretch-It question]. Awesome! Why don't you give that [microphone] to <other GTA> and while you are doing that, how did you think that one [round] went?*
GTA 13: *Actually, I think I got some questions in this time. I asked about their assumptions. My students always forget about the assumptions…*

[Discussion continued]

APPENDIX B

The following is a snippet of dialogue between GTA 6 and the virtual class during their simulator session focused on group dynamics. The context of the session included avatar students working in groups to predict and discuss lightbulbs in three different circuit orientations. In this excerpt, GTA 6 checked in with Sean, Kevin, and Maria about their progress and found the three avatar students might have disagreed with each other causing group dysfunction.

GTA 6:	What, what part did you guys get to?
Sean:	Um, well do you mean like as a group?
GTA 6:	Yes. And not individually cause it's group work.
Sean:	Um, well, yeah, I mean, we made our predictions separately, but we got up to I think, let me look, let me check my work. Okay. Kevin and I got through, up to like right before and then, um, we, I mean, yeah, we kind of, I think we worked through most everything. I mean, Maria was kind of doing the right thing cause, uh, I don't know. She just seemed like she knew she was doing it. Um, she wasn't really great with any of our predictions. And so, she has kind of like went off on her own.
GTA 6:	So, so you guys didn't work with her because she didn't agree with what you guys are predicting? Is that what you're saying?
Sean:	Well, it's not like we didn't want to work with her. It's just that we were making predictions and then she was like saying like that they weren't right. So. Then she, like, we were trying to discuss it with her, but she, I don't know. We just, I don't think we were, we didn't really communicate really well I guess.
GTA 6:	I see. Well, the whole emphasis of the-…
Sean:	She just like set the whole thing up.
GTA 6:	Well, well, the whole purpose of the group work is for you guys to work together and discuss things and you don't have to be right, or you don't have to be wrong, and you just have to, um, see what each other is thinking. And then with the predictions, there's no right [Sean: Yeah.] or wrong. It's just something that you come up with based off of what you feel that should be or off of some type of evidence that you have.
Kevin:	Yeah. I mean, we were saying that, um, you know, to Maria, like we should talk about our predictions and stuff, but I don't know. I mean, I don't want to talk, I don't want to speak for Maria, but I just felt like she just, wasn't, didn't want to hear [pause] [GTA 6: Well, well…] what we had to say. She felt like we were wrong.
GTA 6:	Well, Maria, let me ask you, why did you think that their predictions are wrong? Or why did you feel the need to say that their predictions were wrong?

[1] 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).

[2] President's Council of Advisors on Science and Technology, Report to the president, Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics, Washington DC, https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/pcast-engage-to-excel-final_2-25-12.pdf 2012.

[3] 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).

[4] G. Harris, J. Froman, and J. Surles, The professional development of graduate mathematics teaching assistants, Int. J. Math. Educ. Sci. Technol. **40**, 157 (2009).

[5] T. D. Reeves, L. E. Hake, X. Chen, J. Frederick, K. Rudenga, L. H. Ludlow, and C. M. O'Connor, Does context matter? Convergent and divergent findings in the cross-institutional evaluation of graduate teaching assistant professional development programs, CBE Life Sci. Educ. **17**, ar8 (2018).

[6] D. Lemov, *Teach Like a Champion: 49 Techniques that Put Students on the Path to College* (Jossey-Bass, San Francisco, CA, 2010), pp. 41–47.

[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] 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).

[9] E. C. Goodwin, J. N. Cao, M. Fletcher, J. L. Flaiban, and E. E. Shortlidge, Catching the wave: Are biology graduate students on board with evidence-based teaching?, CBE Life Sci. Educ. **17**, ar43 (2018).

[10] R. M. Goertzen, R. E. Scherr, and A. Elby, Respecting tutorial instructors' beliefs and experiences: A case study of a physics teaching assistant, Phys. Rev. ST Phys. Educ. Res. **6**, 020125 (2010).

[11] 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).

[12] C. Henderson, M. Dancy, and M. Niewiadomska-Bugaj, Use of research-based instructional strategies in introductory physics: Where do faculty leave the innovation-decision process?, Phys. Rev. ST Phys. Educ. Res. **8**, 020104 (2012).

[13] M. Stains and T. Vickrey, Fidelity of implementation: An overlooked yet critical construct to establish effectiveness of evidence-based instructional practices, CBE Life Sci. Educ. **16**, rm1 (2017).

[14] T. M. Andrews, M. J. Leonard, C. A. Colgrove, and S. T. Kalinowski, Active learning not associated with student learning in a random sample of college biology courses, CBE Life Sci. Educ. **10**, 394 (2011).

[15] S. Sandi-Urena and T. Gatlin, Factors contributing to the development of graduate teaching assistant self-image, J. Chem. Educ. **90**, 1303 (2013).

[16] 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).

[17] S. W. Lee, The impact of a pedagogy course on the teaching beliefs of inexperienced graduate teaching assistants, CBE Life Sci. Educ. **18**, ar5 (2019).

[18] G. E. Gardner and M. G. Jones, Pedagogical preparation of the science graduate teaching assistant: Challenges and implications, Sci. Educ. **20**, 31 (2011).

[19] T. Pentecost, L. Langdon, M. Asirvatham, H. Robus, and R. Parson, Graduate teaching assistant training that fosters student-centered instruction and professional development, J. Coll. Sci. Teach. **41**, 68 (2012).

[20] L. B. Wheeler, J. L. Maeng, and B. A. Whitworth, Teaching assistants' perceptions of a training to support an inquiry-based general chemistry laboratory course, Chem. Educ. Res. Pract. **16**, 824 (2015).

[21] E. E. Schussler, Q. Read, G. Marbach-Ad, K. Miller, and M. Ferzli, Preparing biology graduate teaching assistants for their roles as instructors: An assessment of institutional approaches, CBE Life Sci. Educ. **14**, ar31 (2015).

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

[23] D. Doucette, R. Clark, and C. Singh, Professional development combining cognitive apprenticeship and expectancy-value theories improves lab teaching assistants' instructional views and practices, Phys. Rev. Phys. Educ. Res. **16**, 020102 (2020).

[24] 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).

[25] 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).

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

[27] Z. Arsal, Microteaching and pre-service teachers' sense of self-efficacy in teaching, Eur. J. Teach. Educ. **37**, 453 (2014).

[28] M. M. Richard, The impact of micro teaching lessons on teacher professional skills: Some reflections from South African student teachers, Int. J. Teach. Learn. Higher Educ. **10**, 164 (2021).

[29] M. L. Fernández, Learning through microteaching lesson study in teacher preparation, Action Teach. Educ. **26**, 37 (2005).

[30] M. L. Fernandez and M. Robinson, Prospective teachers' perspectives on microteaching lesson study, Education **127**, 203 (2007).

[31] E. Alicea-Muñoz, Transforming the Preparation of Physics Graduate Teaching Assistants, Doctoral dissertation, Georgia Institute of Technology, 2019.

[32] 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).

[33] J. S. Brown, A. Collins, and P. Duguid, Situated cognition and the culture of learning, Educ. Res. **18**, 32 (1989).

[34] K. Kunkler, The role of medical simulation: An overview, Int. J. Med. Robot **2**, 203 (2006).

[35] G. Sroka, L. S. Feldman, M. C. Vassiliou, P. A. Kaneva, R. Fayez, and G. M. Fried, Fundamentals of laparoscopic surgery simulator training to proficiency improves laparoscopic performance in the operating room—a randomized controlled trial, American Journal of orthodontics and oral surgery **199**, 115 (2010).

[36] G. Echegaray, I. Herrera, I. Aguinaga, C. Buchart, and D. Borro, A brain surgery simulator, IEEE Comput. Graphics Appl. **34**, 12 (2014).

[37] R. T. Hays, J. W. Jacobs, C. Prince, and E. Salas, Flight simulator training effectiveness: A meta-analysis, Mil. Psychol. **4**, 63 (1992).

[38] K. Perkins, W. Adams, M. Dubson, N. Finkelstein, S. Reid, C. Wieman, and R. LeMaster, PhET: Interactive simulations for teaching and learning physics, Phys. Teach. **44**, 18 (2006).

[39] C. E. Wieman, W. K. Adams, P. Loeblein, and K. K. Perkins, Teaching physics using PhET simulations, Phys. Teach. **48**, 225 (2010).

[40] C. He and C. Yan, Exploring authenticity of microteaching in pre-service teacher education programmes, Teach. Teach. Educ. **22**, 291 (2011).

[41] A. H. Brown, Simulated classrooms and artificial students: The potential effects of new technologies on teacher education, J. Res. Comput. Educ. **32**, 307 (1999).
[42] L. A. Dieker, C. L. Straub, C. E. Hughes, M. C. Hynes, and S. Hardin, Learning from virtual students, Educ. Leadership **71**, 54 (2014), https://www.learntechlib.org/p/153594/.
[43] A. T. Hayes, C. L. Straub, L. A. Dieker, C. E. Hughes, and M. C. Hynes, Ludic learning: Exploration of TLE TeachLivE™ and effective teacher training, Int. J. Gaming Comput. Mediated Simul. **5**, 20 (2013).
[44] 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).
[45] T. Wan, C. M. Doty, A. A. Geraets, C. A. Nix, E. K. Saitta, and J. J. Chini, Evaluating the impact of a classroom simulator training on graduate teaching assistants' instructional practices and undergraduate student learning, Phys. Rev. Phys. Educ. Res. **17**, 010146 (2021).
[46] A. Geraets, Developing GTA instructional skills: How does a mixed reality teaching simulator Impact GTA Instruction?, Doctoral dissertation, University of Central Florida, 2021.
[47] K. D. Tanner, Structure matters: twenty-one teaching strategies to promote student engagement and cultivate classroom equity, CBE Life Sci. Educ. **12**, 322 (2013).
[48] 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).
[49] 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).
[50] D. F. Feldon, J. Peugh, B. E. Timmerman, M. A. Maher, M. Hurst, D. Strickland, J. A. Gilmore, and C. Stiegelmeyer, Graduate students' teaching experiences improve their methodological research skills, Science **333**, 1037 (2011).
[51] W. Carr and S. Kemmis, *Becoming Critical: Education, Knowledge and Action Research* 2nd ed. (Deakin University Press, Geelong, Victoria, Australia, 1986).
[52] J. M. Case and G. Light, Emerging methodologies in engineering education research, J. Eng. Educ. **100**, 186 (2011).
[53] A. A. Geraets, I. L. Nottolini, C. M. Doty, T. Wan, J. J. Chini, and E. K. 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. Educ. Technol. **30**, 829 (2021).
[54] J. J. Chini and J. W. T. Pond, Comparing traditional and studio courses through FCI gains and losses, *Proceedings of the Physics Education Research, 2014, Minneapolis, MN* (2014), http://doi.org/10.1119/perc.2014.pr.009.
[55] 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/.
[56] E. Etkina and A. Van Heuvelen, 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.
[57] We use the term "high-intensity" to differentiate between other iterations of the simulator training described in our previous work [45,46,53].
[58] N. G. Holmes, C. E. Wieman, and D. Bonn, Teaching critical thinking, Proc. Natl. Acad. Sci. U.S.A. **112**, 11199 (2015).
[59] We used this activity because it focuses on drawing conclusions from evidence (i.e., measurement data) which provides ample opportunities for GTAs to implement Stretch-It questions, and we intend to recruit GTAs who teach with different curricula in future work.
[60] J. Creswell and C. Poth, *Qualitative Inquiry & Research Design* 4th ed. (Sage Publications, Inc, Los Angeles, CA, 2018).
[61] K. Gwet, Kappa statistic is not satisfactory for assessing the extent of agreement between raters, Stat. Methods Inter-Rater Reliab. Assess. **1**, 1 (2002).
[62] K. L. Gwet, *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters* (Advanced Analytics, Gaithersburg, 2014), 4th ed., Vo1. 1, pp. 104–107.
[63] 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).
[64] We observed two GTAs in Fall 2018 teaching the same lesson twice and thus have taken the average of the two observations as one observation.
[65] J. L. Hintze and R. D. Nelson, Violin plots: A box plot-density trace synergism, Am. Stat. **52**, 181 (1998).
[66] RStudio Team, *RStudio: Integrated Development for R. RStudio* (PBC, Boston, MA 2022). http://www.rstudio.com/.
[67] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Spring-Verlag, New York, 2016).
[68] Periscope video lessons available to physics instructors at physport.org/periscope.
[69] J. D. Gaffney, A. L. H. Gaffney, and R. J. Beichner, Do they see it coming? Using expectancy violation to gauge the success of pedagogical reforms, Phys. Rev. ST Phys. Educ. Res. **6**, 010102 (2010).
[70] E. K. Saitta, M. Wilcox, W. D. James, and J. J. Chini, The views of GTAs impacted by cross-tiered professional development: Messages intended and received, Int. J. Res. Undergrad. Math. Educ. **6**, 421 (2020).