Implementation of an investigative science learning environment based laboratory course taught by novice instructors

Joshua Rutberg, Diane Jammula, and Sheehan Ahmed
Department of Physics, Rutgers University—Newark, Newark, New Jersey 07103, USA

(Received 26 May 2023; accepted 27 September 2023; published 30 October 2023)

[This paper is part of the Focused Collection on Instructional labs: Improving traditions and new directions.] This study investigated the development of four new novice graduate teaching assistants (GTAs) and one nontenure track (NTT) faculty member assigned to teach an introductory physics laboratory course based on the investigative science learning environment (ISLE) approach. We describe the training that these instructors received and the progress that they made over their first semester of instruction. We used classroom observations scored using the Danielson Framework and the Reform Teaching Observation Protocol (RTOP) to assess each instructor's teaching multiple times throughout the semester. All four GTAs demonstrated improvement of various degrees while the NTT did not. Classroom observations and debriefs improved the GTAs' teaching and were used to tailor weekly training meetings to their needs. Our results suggest that new instructor training should feature specific, tangible teaching actions which new instructors can implement easily in their own classrooms. When these actions directly address issues that instructors have had, as indicated by conversations with the instructors and classroom observations, they were more likely to be adopted and have a noticeable effect on the quality of teaching.

DOI: 10.1103/PhysRevPhysEducRes.19.020153

I. INTRODUCTION

Recent calls for reform in physics education have drawn attention to the need for many kinds of change from curriculum to classroom technologies. But we also need qualified and knowledgeable instructors to teach these classes. An excellent curriculum does little without an instructor who knows how to implement it. Despite this importance, little attention is paid at the university level to the way new instructors enter the profession. New instructors, including graduate teaching assistants (GTAs), are regularly assigned teaching loads with little or no training to prepare them to teach [1,2]. Whether this is because of outdated assumptions that only content knowledge is required to teach, a lack of concern about the quality of instruction these new instructors provide, or institutional inertia, the end result is that new instructors often have little choice but to recreate the traditional methods by which they were taught [3].

The study described in this paper took place during our reform of the introductory physics program at Rutgers University-Newark. This introductory program consists of two separate lecture courses—one algebra-based and the other calculus-based—and a single-shared laboratory course, all of which are taught using the Investigative Science Learning Environment (ISLE) approach [4,5]. All of the large enrollment "lecture" courses and two sections of the laboratory course were taught by members of the reform team, with the remaining sections of the lab taught by graduate teaching assistants (GTAs), part-time lecturers (PTLs), and a nontenure track (NTT) faculty member. The four GTAs teaching the course during this study were all in their first year, new to teaching and to the institution. While the NTT had 12 years of teaching experience at other institutions, this was his first experience with student-centered teaching.

Teaching a traditional course with no preparation is a difficult task; teaching a student-centered class with no preparation is significantly more challenging [6,7]. The ISLE-based laboratory course our instructors were assigned to teach was student centered and placed a heavy emphasis on active learning. Teaching this course required pedagogical skills such as facilitating group and class discussions, active listening, managing group work, and being able to help students design their own experiments without giving them a specific procedure to follow [5]. It also required an understanding of the scientific practice central to the ISLE approach. This style of teaching is difficult even for instructors with years of traditional teaching experience.

Compounding this difficulty were the structural limitations placed on the amount of support we could provide

TABLE I. A sample of physics reforms with laboratory components and their broadly defined goals.

Lab reform	Constructing knowledge	Scientific modeling	Designing experiments	Developing skills	Analyzing data	Communicating ideas
SQILabs [15]				✓		
Modeling instruction [16]	✓	✓				
SCALE-UP [17]			✓		✓	
DL-SCL [19]	✓	✓	✓			
Scientific Community Labs [20]			✓	✓	✓	
The ISLE approach [5]	✓	✓	✓	✓	✓	✓

these new instructors. At the time of this study, there was no formal mentoring program for new instructors nor was much time available for pedagogical training.

Nevertheless, we observed how these instructors performed over the course of their first semester teaching with the support we could offer. In doing so we wanted to answer the following research questions:

- What aspects of teaching an ISLE-based course are most difficult for novice instructors?
- How do the teaching practices of novice instructors change over their first semester of teaching an ISLE-based lab course?
- Which training practices have the largest effect on novice instructors' teaching practices?

II. REFORMED PHYSICS LABS

A. Other physics laboratory reform efforts

Calls to reform introductory physics labs have been ongoing for decades [8,9]. While many physics professors have long had an intuitive feeling that laboratory experiences are an essential part of learning physics [10], there has never been widespread consensus about what the goals of laboratory instruction should be and its benefits have never been clearly demonstrated [11]. Traditional labs do not appear to help students better understand physics content [12] and usually leave students with less expert beliefs about the experimental nature of physics [13].

In 2014, the American Association of Physics Teachers laid out a set of goals and recommendations for reforming introductory physics labs [14]. They recommended six potential goals of laboratory instruction: constructing physics knowledge, developing and testing models, designing experiments, developing technical skills, analyzing and visualizing data, and communicating physics knowledge and experimental results. These goals align with those of many of the programs meant to reform introductory physics labs.

For instance, the Structured Quantitative Inquiry Labs (SQILabs) format [15] aims to teach students to develop the technical skill of using measures of uncertainty to compare datasets to a theoretical model and make decisions about whether to accept, modify, or reject the model based on those comparisons. Modeling Instruction [16] has an experimental component focused on constructing physics knowledge by developing and deploying models. Student-Centered Activities for Large Enrollment Undergraduate Programs (SCALE-UP) [17] has a laboratory component that emphasizes designing experiments and analyzing data. Table I contains a sample of reformed introductory physics programs with laboratory components and a general list of their implicit or explicit goals. A more thorough review of reform-oriented lab courses found that 46% of first-year lab courses surveyed identified developing lab skills and reinforcing physics concepts as their primary goals, 20% aimed only to reinforce concepts, and 34% aimed to only teach experimental skills [18]. While the goals of these reformed laboratory curricula may be different, they share many similarities with regard to how they are taught. All are heavily student focused, as opposed to traditional labs where students are expected to follow a predesigned procedure with a known outcome. In such a traditional course, the instructor's primary role is that of a supervisor and troubleshooter. Reformed labs, by contrast, are almost always active and student centered. They require an entirely separate skillset in order to teach.

B. The ISLE approach

Rather than a set of curricular materials, the ISLE approach is a framework for designing and enacting physics instruction where students construct physics concepts by going through the same logical processes as practicing physicists. Instead of choosing between developing skills—experimental or reasoning—or teaching physics content, the ISLE approach holds that these two goals are inseparable from one another.

In the ISLE approach, students use hypothetico-deductive reasoning [21] to construct "new" physics ideas for themselves rather than being told what to believe by professors or a textbook. The experimental skills by which students are assessed, embodied in the Scientific Abilities Rubrics [22], are the concrete steps into which this complex process can be reduced. In this way, physics content is not "verified" or "reinforced" in an ISLE-based course but developed by students from experiments and through specific reasoning approaches (inductive and hypothetico-deductive reasoning). The ideas developed by the students in the lab are discussed and expanded in "large room meetings" (traditionally called lectures, but in the ISLE approach, they are also interactive). The lab work and the "lecture" work are closely intertwined and not independent.

The chain of reasoning used by students in the ISLE approach begins with an anomalous observation that they cannot yet explain or a question to which they do not know the answer. This serves as a 'need to know' which provides the motivation for learning new things; students are not simply learning kinematics because that is what the curriculum says they will learn; they are trying to understand how Google Maps can estimate how long it will take them to drive home after class.

Once this need to know has been established, students perform experiments intended to collect more data, from which patterns can be identified and hypotheses generated. We call these experiments—experiments done for the purpose of learning enough about a phenomenon to produce one or more hypotheses—*Observational experiments*. After students have developed one or more plausible hypotheses, they must design and conduct experiments for the purpose of testing these hypotheses. We call these experiments *testing experiments*.

A testing experiment is an experiment whose outcome can be predicted by the hypothesis. If multiple hypotheses are being tested, then they should lead to distinct predictions about the outcome. Once a prediction has been made, the experiment can be conducted and the outcome recorded. Students then compare that outcome to the prediction made by each hypothesis and decide whether the experiment supports or refutes that hypothesis.

When professional scientists are performing authentic research, it often takes multiple different, independent testing experiments before a new hypothesis will be accepted, but because of the time constraints of the typical introductory physics course, this is not always possible to replicate with students. Going through this process themselves gives students a better appreciation not just of the physics knowledge they have developed, but of the process by which all new knowledge is constructed. Lawson has done an admirable job outlining how this way of reasoning was used in many historically significant scientific discoveries [23], but this type of hypothetico-deductive reasoning is not limited to scientific disciplines. It is, for example, the process by which doctors diagnose their patients.

Finally, once an idea has been tested and accepted, it can be used to answer new questions and solve problems. This can be done in *application experiments*. Figure 1 provides an illustration of how this process plays out in the classroom, adapted from the one found in Etkina *et al.* [5].

Beyond this emphasis on hypothetico-deductive reasoning, the ISLE approach stresses several social practices that are just as much a part of scientific practice as

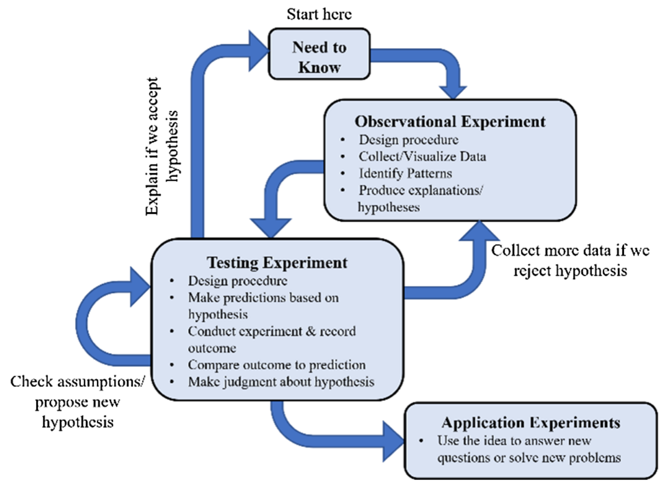


FIG. 1. The ISLE approach to scientific reasoning.

experimentation. While none of these are necessarily unique to the ISLE approach, all of them are considered essential to ISLE-based instruction.

- An emphasis on fostering a 'growth mindset' [24], often by providing opportunities for students to correct and resubmit past work for credit and reassessing the same skills across multiple assignments.
- Building a "community of practice" [25] where all students are encouraged and expected to participate in the development of new ideas collaboratively.
- An emphasis on using and translating between multiple physics representations [26].

C. Sample ISLE-based lab session

To illustrate the way that the ISLE approach was used in our course, we will present the instructional activities from a sample lab lesson. The purpose of this lesson is for students to develop and test the idea of Newton's third law. The "large room meeting" that follows builds and expands on the knowledge that the students develop in the lab.

1. Need to know

To begin this lesson, the instructor shows students what happens when someone sitting on a rolling chair pushes another person sitting in a rolling chair. They should see that while the person being pushed rolls forwards, the person who did the pushing rolls backward. From this, they should infer that each person was exerting a force on the other and that those forces are in opposite directions. However, students do not yet know how the magnitudes of those two forces compare to one another. To figure that out, they will need to collect data.

2. Observational experiment

In our lab, we use Vernier® force sensors connected to Lab Quest devices, so students can record graphs showing

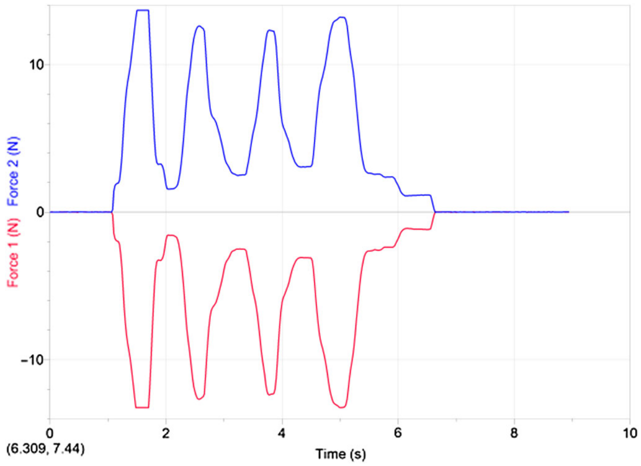


FIG. 2. Sample data from the observational experiment obtained by pushing two force sensors together. The two readings are always equal and opposite.

the force exerted over time. With two force sensors connected, students can have them push and pull on each other in various ways and record the resulting graphs. The details of how they perform this experiment are up to them, and they are encouraged to try a number of different experiments such as dragging one force sensor along the table using the other, holding one force sensor in place and pushing or pulling it with the other, or pressing the two force sensors into one another.

Regardless of what they do with the sensors, students will get a graph similar to the one in Fig. 2. By sharing their data with other students who performed different experiments (pulling instead of pushing, for example), all students should identify the pattern that "the readings on the two forces sensors are equal and opposite no matter what is being done with the force sensors." An explanation for this could be that *when two objects interact, the forces they exert on one another are always equal and opposite*. This explanation is a hypothesis that can be tested.

3. Testing experiment

A good testing experiment should try to disprove the hypothesis, so students need to think of when the pattern they came up with is most likely to break. Many students will think that the forces in the previous experiment were only equal in magnitude because identical force sensors were used. A good way to test this idea is to attach the force sensors to carts of different masses. While many students may personally think that the more massive cart will exert a larger force when the two carts collide, the hypothesis that they came up with in the previous experiment predicts that the readings on the two sensors should still be equal and opposite. Figure 3 shows the setup students would use to conduct this experiment and an example of the data they might obtain from this experiment.

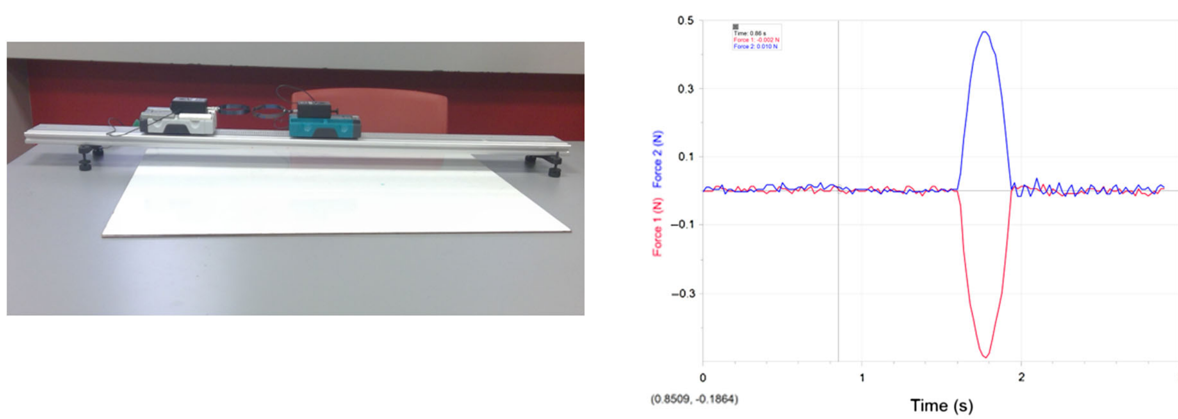
Since the outcome of all these experiments matches the prediction made by the hypothesis, students are unable to reject this idea and come to accept that it is true.

Notice that the students construct Newton's third law in this lab as well as develop experimental and reasoning skills. The "large room meeting" that follows engages the students in solving problems using Newton's third law.

III. TRAINING NOVICE LAB INSTRUCTORS

A. Difficulties faced by novice instructors

Previous research into how new instructors enter the profession paints a somewhat bleak picture. A landmark study about the experiences of new college instructors found most new instructors receive little to no pedagogical support from their more experienced colleagues [27]. More recent studies [28] suggest this is still the norm despite studies suggesting that collegial support is key to the success of new faculty members [29,30]. This problem affects more than just new instructors; one study found that a third of all faculty members surveyed felt like they received no support from their colleagues or university when it came to their classroom instruction [31]. When



(a) (b)

FIG. 3. (a) the experimental setup used to test the idea that when two objects exert forces on one another those forces will be equal and opposite and (b) a sample outcome from this experiment that supports this idea.

faculty do receive support from their colleagues, it is most often by having course materials such as lecture slides passed down to them rather than conversations about teaching or pedagogical advice [32].

This lack of support can have particularly negative effects on how new faculty teach. Time constraints and lack of guidance lead many to simply copy the pedagogical style of their own instructors, even when they do not believe those methods are effective [3]. This is especially true of doctoral students asked to take on teaching roles who are expected to prepare for and teach classes alongside their own coursework and participate in research, often with little advanced warning [32].

While some universities are beginning to offer teacher development programs and pedagogical training for new or future faculty members [33,34], these programs are the exception and not the rule. Other programs such as the Workshop for New Physics and Astronomy Faculty [1] are designed to expose as many beginning physics professors as possible to research-based instructional strategies, but one-time exposure to these strategies is not sufficient to ensure accurate and sustained implementation of these strategies [35].

Without support, new instructors have no option but to replicate the style of teaching of their own instructors, perpetuating outdated and ineffective modes of instruction such as passive lecturing. Even if they are aware of the evidence that such modes of instruction have little benefit for students [36,37], they often lack the time or knowledge to implement research-based instructional strategies for themselves.

Untrained faculty often possess simplistic views of what makes for effective teaching, for example, that "good teaching equals clear, knowledgeable, and, possibly, inspiring lectures" [27], (p. 157). This simplistic belief about effective teaching can lead new instructors to overestimate their teaching ability and show little interest in improvement. For example, in one study, it took three semesters of receiving unsatisfactory student evaluations before the instructors would acknowledge the problems with their teaching [27].

B. Knowledge required to teach reformed labs

Facilitating a traditional lab course requires very little pedagogical skill. When students are expected to follow predesigned procedures, whose outcomes are known in advance to both students and instructors, the primary job of the instructor is supervision and troubleshooting, not actual teaching. The most teaching an instructor in a traditional lab course may do is present a prepared lecture to students about the theory behind the lab at the beginning of class.

Teaching a lab where students actually learn physics content, scientific skills, or both is significantly more complex. It requires a synthesis of content knowledge, knowledge of theories of teaching and learning, and pedagogical content knowledge [38] or content knowledge for teaching [39]. Most of these are knowledge bases that instructors are unlikely to develop during their own undergraduate or master's education.

Pedagogical knowledge, or knowledge of theories of teaching and learning, is a broad base of subject-nonspecific teaching strategies and learning theories. To be an effective teacher, one must understand how the physical process of learning occurs in the brain [40] as well as theories of learning that attempt to explain how people make sense of new information such as constructivism [41]. For example, a reformed lab instructor must understand that knowledge cannot be transferred into students' heads, but students must identify, play with, and test their ideas to build new connections in their brains.

Pedagogical knowledge also includes practical knowledge of the concrete skills necessary for the daily maintenance of an effective classroom, such as how to facilitate discussions, manage time, and keep students engaged. In a reformed lab, instructors need to help students communicate ideas with each other while also budgeting time appropriately.

Pedagogical content knowledge, or content knowledge for teaching [39], is a domain of knowledge specific to teachers of certain subjects. It is not merely knowing how to teach or knowing a subject but knowing how to teach a subject. Magnusson *et al.*, [42] define pedagogical content knowledge as "a teacher's understanding of how to help students understand specific subject matter" (p. 96) and identify five key components of pedagogical content knowledge for science: orientation toward teaching science, knowledge of science curriculum, knowledge of students' understanding of science, knowledge of assessment in science, and knowledge of instructional strategies.

Beyond knowing what content students need to learn and what curricular materials are available, a reformed lab instructor needs to know how research-based materials are meant to be used. Teaching a reformed lab successfully entails understanding not just what students will be expected to do, but why they are doing it. Simply having educational resources is no guarantee that an instructor will understand the purpose of those materials and how to use them.

For example, Robertson [43] found that novices enrolled in her learning assistant program were unable to identify the instructional goals behind activities from the tutorials in physics [44] materials used in their courses. Henderson and Dancy [35] found that physics faculty members who attempted to adopt research-based instructional strategies often modified these strategies in ways that, without their knowledge, fundamentally compromised the benefits of those particular strategies. Because of the open-ended nature of reformed labs, instructors need to know how students learn particular content in order to best decide where to focus discussion when teaching a lesson.

Knowledge of students' understanding of science matters because an effective teacher needs to be able to anticipate where students are likely to struggle and have on-hand questions or activities meant to guide students through those difficulties. For example, in the lesson described above where students develop and test Newton's third law, instructors need to know the common conception that a bigger object exerts more force than a smaller object when they interact. Instructors need strategies to help students work through this idea.

There are other pedagogical skills and knowledge that are specific to only certain types of lab reforms. For instance, any course where students are expected to develop their own experimental procedures is significantly more difficult to teach. In order to teach a course like this effectively, an instructor must be able to quickly evaluate an experimental procedure they have never seen before and decide whether or not it is practical and will accomplish the goals of the lesson and experiment.

C. Knowledge required to teach with the ISLE approach

In addition to the pedagogical skills required of many other student-centered lab courses, teaching an ISLE-based course requires a specific set of skills and knowledge that many instructors new to this approach are unlikely to have. Foremost among these is knowledge of the ISLE approach itself. For the lab described above to be taught successfully, an instructor must know the difference between observational experiments and testing experiments and must be able to communicate that difference to students. This is nontrivial. While the primary difference between these types of experiments is the objective of the experiment—observational experiments are done to collect data, identify patterns, and formulate hypotheses while testing experiments are done to test existing hypotheses—many students and new instructors believe the difference between them is primarily in their procedures. They might incorrectly believe that observational experiments must be passive and any experiment where quantities are being directly controlled must be a testing experiment, regardless of the purpose behind the experiment. If the instructor does not have the knowledge and ability to guide students through this process, students are unlikely to engage in scientific reasoning and will instead simply go through the motions of conducting experiments and collecting data.

For students to understand how to perform a testing experiment, the instructor must understand and be able to communicate the difference between a hypothesis and prediction, concepts that are often conflated in other approaches to experimentation [45].

This ISLE-specific knowledge can take time for instructors to learn. In a master's program specifically designed to train preservice teachers to teach using the ISLE approach [46], it can take up to 2 years before beginning teachers begin to feel comfortable with this approach [47].

D. Training of instructors in reformed labs

Despite the abundance of research into introductory lab reform, there has been little written about how instructors are trained to teach these courses. This is surprising considering how much more difficult these courses are to teach than traditional lab courses.

Several others have documented the ways they train their GTAs [48–51]. However, few studies evaluate the effectiveness of that training. Most often the impact of GTA training programs is assessed by interviews with the GTAs and measures of their confidence or by interviews with faculty who work alongside the GTAs. There are very few reports of the classroom practices of GTAs and how they are affected by training. We aim to fill this gap by recording exactly how the instructional practices of our instructors change over the course of their first semester of teaching.

IV. STUDY SETTING

This study took place within the broader project of reforming all aspects of the introductory physics courses at Rutgers University-Newark using the ISLE approach. Here we will discuss the details of that reform which are relevant to the study and outline the training we provided our instructors.

A. Institutional setting

Rutgers University-Newark is a diverse, urban institution. The student population is 31% Hispanic, 20% White, 18% Black, 18% Asian, and 13% from different backgrounds. 59% of students are Pell-eligible, meaning they have exceptional financial need. These demographics are important because the vast majority of physics education research (PER) has been conducted at institutions that do not resemble ours; PER is typically conducted at affluent, predominantly white institutions [52]. For this reason, investigations into physics education occurring at institutions like ours fill a gap in existing research.

Another unique consideration of Rutgers University-Newark is that this study was conducted when the first cohort of our revamped graduate program began. This meant that we could change our introductory physics program without the challenge of institutional memory. However, lines of communication or guidance that may exist in more established programs were still being developed. The training provided to the instructors over the course of this study was the only training these instructors received in how to teach.

B. Course context

During the 2019–2020 academic year, three courses were revised using the ISLE approach: an algebra-based introductory physics course ($N = 300$), a calculus-based introductory physics course ($N = 60$), and a corresponding one-credit laboratory course ($N = 350$). The laboratory course serves as a companion to both the algebra-based and calculus-based courses, a restriction which predated our reform and which we were unable to change. Since the two courses share a lab, they must follow the same sequence and timing to keep the lab in synch with both courses.

All three of these courses were completely reformed using the ISLE approach. For the laboratory course, a curriculum consisting of ten labs was developed based on the lab courses developed at Rutgers University-New Brunswick [53].

Ten sections of the laboratory course were offered during the Fall 2019 semester. Of these, two were taught by the head of the reform team, one each was taught by four new graduate student teaching assistants (GTAs), two were taught by adjunct instructors, and two were taught by a new nontenure track (NTT) faculty member. This study focuses on the four GTAs, all of whom were first-year doctoral students with no formal prior teaching experience. We also followed the NTT, with 12 years of traditional teaching experience at other institutions, since he served as an interesting counterexample to the trends observed in the GTAs.

C. Reformed labs

Earlier, we provided a sample of a reformed lab activity from our curriculum, but here we will outline the ways in which our reformed lab course differs from traditional lab courses—including the course taught at our university prior to our reform. We will focus on the differences that directly impact the way the course is meant to be taught.

Besides the close relationship with the lecture course, the biggest difference between the reformed curriculum and a traditional course is the structure of each lab session. In the old course, each lab consisted of a single experiment which groups of students were expected to complete independently from all other groups before leaving. No discussion between groups was expected. The purpose of the lab was for students to conduct a specific predetermined experiment, get a specific known result, and confirm a law or concept previously taught to them in a lecture.

In the reformed curriculum, each lab consists of multiple smaller experiments that serve different purposes in the process of scientific inquiry [54] and create a cohesive narrative about the ideas students are meant to develop. This might include an observational experiment to devise a new hypothesis and a testing experiment to test it, as in the example given above, or it might include experiments to first test and then apply an idea that students constructed in the nonlab portion of the course. The instructor is expected to facilitate a “board meeting” [55] before and after each experiment where groups can share their findings with one another and establish a learning community [56].

Another important difference between the two curricula is the extent of the guidance given to students. In the old curriculum, students were told which measurements to take, how those measurements should be taken, and how their data should be analyzed. Students simply followed instructions to see what they were “supposed to” see regardless of whether they understood anything they did. The reformed curriculum, on the other hand, regularly asks students to design their own experiments because the intentional nature of the curriculum means that the process through which the students work is just as, if not more, important than the result of a particular experiment. In addition to developing new understandings of the physics content of the lessons as a part of their learning of physics, our goal is for students to develop an understanding of how to design effective experimental procedures.

V. PROFESSIONAL DEVELOPMENT

This section describes the professional development with which we were able to provide our instructors to prepare them to teach this course.

A. Presemester workshop

All introductory physics instructors, GTAs, and undergraduate learning assistants teaching our introductory courses are required to attend a 5-h presemester workshop. During this workshop, they go through an ISLE-based learning cycle and discuss the essential aspects of the ISLE approach. For most of the participants, this is their first exposure to the ISLE approach. The purpose of this workshop is to introduce them to the philosophy behind the ISLE approach and help them begin to make sense of the way we expect them to teach.

B. Weekly training meetings

The primary source of professional development came from weekly training meetings. These meetings were 80 min long, shorter than the 110-min lab periods. Prior to the reform, these meetings were meant to inform instructors of the materials students would be using and the procedures they were expected to follow. Given the increased pedagogical skill necessary to teach the reformed curriculum effectively and our GTA’s lack of teaching experience, we decided that these meetings needed to include pedagogical training in addition to familiarizing the instructors with the new course materials. Furthermore, considering the research on new instructors’ lack of support [28], we tried to make these meetings a place where instructors could get feedback on specific problems they faced in their classrooms and foster the feeling that they were part of a community of teachers. Despite the vastly increased scope of these weekly meetings, the length of the meetings remained unchanged for logistical reasons beyond our control.

The bulk of each meeting consisted of a member of the reform team modeling the role of instructor for that week's lab while the other instructors worked as if they were students. We modeled many of the pedagogical strategies we expected them to use in their own classes including questioning techniques, strategies for getting students to work and think collaboratively, and ways to facilitate whole class discussions. Unfortunately, class discussions were hard to simulate realistically with the small number of participants and the fact that some GTAs struggled to take on the role of undergraduate students. This may be because of their more advanced level of physics content knowledge or their insecurity about physics content knowledge and knowledge of science practices.

After each experiment, we paused for discussions about the difficulties they anticipated students having and ways they as instructors might help. We also needed to discuss the purpose of each experiment and what understanding students should take away from each experiment. This proved to be a stumbling point for the instructors used to traditional labs, where the purpose of an experiment is often simply to collect data. In contrast, experiments in an ISLE-based course have different purposes which must be understood by both instructors and students. It is not enough to understand what data must be collected; the students must understand why they are collecting those data. The purpose of a testing experiment is not to make a prediction, it is to test a hypothesis by comparing that prediction with the outcome. This is a significant departure from traditional laboratory instruction and thus the purposes of each experiment were always emphasized in training meetings.

Another important aspect of these training sessions was providing the instructors with time to voice problems they were having and receive feedback. Some of these problems were common to many instructors, such as having difficulty pacing their classes, while others were more specific, such as dealing with students habitually arriving late to class.

1. Pedagogical aspect of training meetings

Because of the limited amount of time available for these meetings, most of our time consisted of having instructors conducting the experiments so that instructors would have first-hand knowledge of what students would be expected to do in their classes. However, as mentioned earlier, we recognized that because of the lack of experience of the instructors and the complexity of the teaching required for these labs, we knew that instructors also needed some guidance in how to teach effectively. Ideally, this would occur in a pedagogy course new GTAs are required to take, but since that was not possible for us, we chose to put as much pedagogical training into our training meetings as possible.

We did not have a particular set of topics in mind for the pedagogical component of these training meetings. Rather, we chose to emphasize skills and ideas based on the problems we saw while observing the instructors. For instance, observing instructors asking only closed questions during their classes prompted us to discuss the difference between open and closed questions in our next training meeting and offering strategies for how instructors can ensure they are asking questions which provoke thought and discussion rather than merely checking for recognition. Topics we discussed included: how to use the whiteboards available in the lab room effectively, various strategies for managing class time, strategies for asking questions and giving students the opportunity to think before responding, how to use ISLE-specific vocabulary, how to provide feedback to students, and the importance of making sure students understand their goals as well as how to effectively communicate goals to students.

C. Observations and coaching

Another source of training for the new instructors came in the form of classroom observations. Each new instructor was observed between 4 and 6 times over the course of the semester. While these observations were initially intended purely for data collection, after the first few observations, they adopted the secondary purpose of coaching and providing real-time feedback. Interference was kept to a minimum to avoid usurping the instructor's authority, but instructors were notified of problems we observed, given suggestions for how to proceed, or received feedback on how the class was going and where improvements could be made. Often the instructors would be reminded of things that had been discussed during that week's training meeting or be given advice on things such as when to move on from one experiment to the next or how to recapture the attention of students.

After each observation, the observer debriefed with the instructor. The instructor shared their thoughts about how the lesson went, listened to the observer's notes, and left with a series of specific, tangible things to work on for future lessons. All GTAs and the NTT participated in these observations and debriefs.

D. Additional voluntary opportunities

While the weekly meetings and class observations were the only mandatory sources of training, two other sources of training existed that instructors could use if they wanted. First was the opportunity to observe the classes taught by a member of the reform team. This instructor was the one who wrote the new lab curriculum, ran the weekly training meetings, and was an expert in the ISLE approach. The sections taught by this expert instructor were always scheduled to be the first classes taught each week so any instructor who wanted to observe how the class was taught with real students would be able to do so before teaching their own class. This open invitation was available for all instructors throughout the semester.

Second, instructors were asked to write and share reflections on each class after teaching every week. Both the act of reflecting and the opportunity to read the reflections of others were intended to help the new instructors. Each week the member of the reform team teaching the first sections of the course modeled this reflection for all instructors to see and expected each instructor would follow suit. However, there was no penalty for not submitting reflections. As such, this was seen as an optional part of the training by the instructors.

VI. DATA COLLECTION AND ANALYSIS

The primary source of data used was classroom observations of the instructors. All four GTAs and the NTT were observed. Two of the GTAs and the NTT were observed 4 times and the other two GTAs were observed 5 times.

Each observation lasted the entire 110-min class period. The observer took notes throughout the observation and recorded audio whenever the instructor spoke. Within a day of each observation, the instructor was debriefed, the notes were shared with them, and areas of potential improvement were discussed.

A. Observing ISLE-specific behaviors

To date, no observation protocol has been developed to measure the fidelity of the implementation of the ISLE approach in the classroom. However, since the observers were experts in the ISLE approach, we were able to specifically look for aspects of the instructors' classroom practices that are considered essential to the success of the ISLE approach and record them in our observation notes. Such practices included whether students were presented with a "need to know" which provided the motivation for the experiments, whether students were able to correctly identify each experiment as an observational, testing, or application experiment and explain how they were different from one another, whether students were made aware of the reasons they were conducting specific experiments, the extent to which instructors allowed students to come up with their own experimental procedures, and whether ISLE-specific terminology such as the difference between a hypothesis and prediction was used correctly throughout the lesson.

Rather than simply observing whether these practices were included in a lesson, we also watched students as they worked. In our observation notes, we tried to record not just what our instructors did, but whether their approaches and strategies had the intended effects on students.

B. Quantifying the quality of instruction

Beyond recording these ISLE-specific teaching practices, we wanted a way to analyze the overall performance of the instructors we observed. To do this, we developed an observation protocol using two existing instruments: the Reformed Teaching Observation Protocol (RTOP) [57] and the Danielson framework [58]. Both of these instruments were designed to measure student-centered, active learning and have been widely used to assess the classroom practices of instructors. Other instruments such as the Teaching Practices Inventory [59] were considered but ultimately rejected because they were more holistic rather than focusing specifically on the act of classroom teaching. The GTAs in this study had no control over designing the instructional activities or choosing assessments for students so instruments that focus solely on what an instructor does in the classroom were preferable to those that encompass all elements of teaching. For this reason, only domains 2 and 3 of the Danielson framework were used.

While both RTOP and the Danielson framework are designed to measure active learning, the two instruments are structurally different from one another. RTOP consists of statements that describe an active learning environment and a Likert scale rating of how descriptive of the lesson that statement is. The Danielson framework takes the form of a rubric with clear descriptors for each level. Since they are both designed to measure the same construct in different ways, we felt including both would provide some additional robustness to the observation scores since we were unable to use multiple independent observers. Table II below demonstrates both the similarities and differences between these two instruments. The complete observation protocol can be found in Supplemental Material [60].

The lesson was rated immediately after class ended using the observation protocol. These ratings were shared with the instructors alongside the notes taken during the observation.

RTOP is broken into three sections, called RTOP A, RTOP B, and RTOP C from this point onward. Similarly, the Danielson framework is broken into 4 domains, of which only domains 2 and 3 pertain to classroom instruction. This left us with 5 domains over which each instructor was rated for each observation. Table III lists the aspects of teaching each domain is meant to describe.

C. Analysis

We analyzed the qualitative and quantitative data collected from our observations independently of one another in a convergent mixed methods approach [61]. To assign a quantitative value to each observation, equal weight was given to each of the five observed domains, and scores in each were normalized to account for the difference in scale between RTOP and Danielson. Simply comparing raw numbers would heavily bias RTOP scores since each RTOP domain consisted of more items than each Danielson domain. To account for this, each domain was scaled to account for one fifth of the total rating, with the highest possible rating given a value of 1.

This gave us a quantitative value to describe the quality of instruction in any observation, allowing us to compare

TABLE II. A comparison of similar aspects of teaching assessed by RTOP and the Danielson framework.

RTOP			
B7: The lesson promoted strongly coherent conceptual understanding			
B8: The teacher has a solid grasp of the subject matter content inherent in the lesson.			
B10: Connections with other content disciplines or real world phenomena were explored and valued.			
C19: Student questions and comments often determined the focus and direction of classroom discourse.			
Danielson framework			
3a: Communicating with students			
Unsatisfactory	Basic	Proficient	Distinguished
The instructional purpose of the lesson is unclear to students, and the directions and procedures are confusing. The teacher's explanation of the content contains major errors and does not include any explanation of strategies students might use. The teacher's spoken or written language contains errors of grammar or syntax. The teacher's academic vocabulary is inappropriate, vague, or used incorrectly, leaving students confused.	The teacher's attempt to explain the instructional purpose has only limited success, and/or directions and procedures must be clarified after initial student confusion. The teacher's explanation of the content may contain minor errors; some portions are clear, others difficult to follow. The teacher's explanation does not invite students to engage intellectually or to understand strategies they might find useful when working independently. The teacher's spoken language is correct by using vocabulary that is either limited or not fully appropriate to the students' ages or background. The teacher rarely takes opportunities to explain academic vocabulary.	The instructional purpose of the lesson is clearly communicated to students, including where it is situated within broader learning; directions and procedures are explained clearly and may be modeled. The teacher's explanation of content is scaffolded, clear, and accurate and connects with students' knowledge and experience. During the explanation of content, the teacher focuses, as appropriate, on strategies students can use when working independently and invites student intellectual engagement. The teacher's spoken and written language is clear and correct and is suitable to students' ages and interests. The teacher's use of academic vocabulary is precise and serves to extend student understanding.	The teacher links the instructional purpose of the lesson to the larger curriculum; the direction and procedures are clear and anticipate possible student misunderstanding. The teacher's explanation of content is thorough and clear, developing conceptual understanding through clear scaffolding and connecting with students' interests. Students contribute to extending the content by explaining concepts to their classmates and suggesting strategies that might be used. The teacher's spoken and written language is expressive, and the teacher finds opportunities to extend students' vocabularies, both within the discipline and for more general use. Students contribute to the correct use of academic vocabulary.

instructors to one another and to look for changes in each instructor's performance over the semester in a statistically significant way.

While the quantitative data could say whether the instructors improved and compare growth among the instructors, we needed qualitative data to describe precisely how the instructors differed from one another and which aspects of their instruction changed. We analyzed the notes taken during the observation using an emergent coding process to construct a portrait of each instructor's teaching over the course of the semester. From this, we were able to identify areas of improvement and patterns across multiple GTAs. These patterns could then be compared to our training practices to see which elements of the training were adopted by our instructors.

TABLE III. Description of each domain observed by RTOP and Danielson.

Domain	Description
RTOP A	Lesson design and implementation
RTOP B	Content
RTOP C	Classroom culture
Danielson 2	The classroom environment
Danielson 3	Instruction

VII. QUANTITATIVE FINDINGS

We report our quantitative findings before our qualitative findings because these data tell us whether the instructors changed significantly over the course of this study while the qualitative data show us what aspects of their instruction changed.

A. Measurements of instructor performance

Figure 4 shows the breakdown of scores given to each instructor over the semester. Observation scores for all four GTAs increased over the course of the semester according

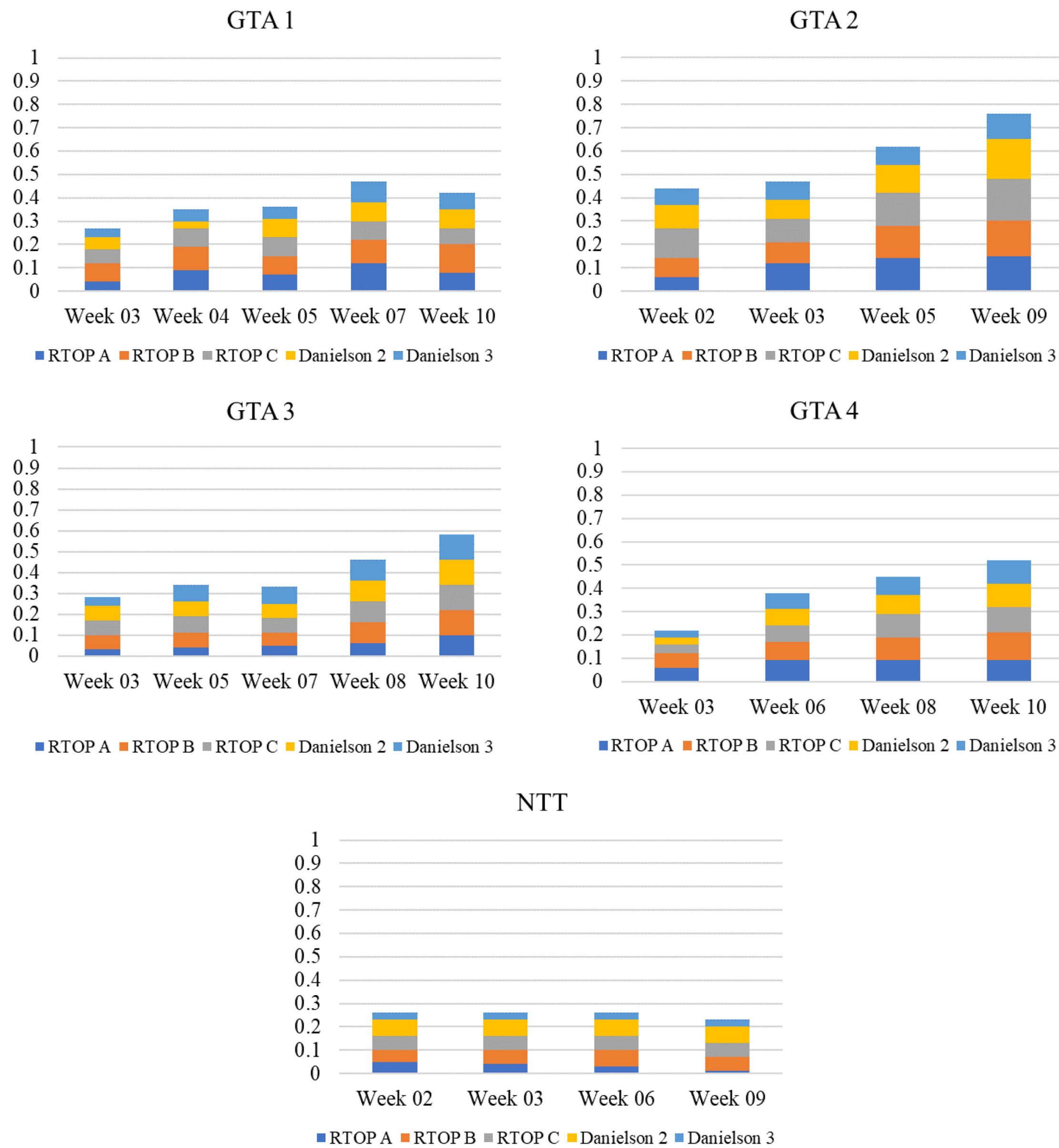


FIG. 4. Observation scores for each instructor broken down by domain.

to both RTOP and Danielson. The NTT, in contrast, showed no change by either measure.

We used regression analysis to determine whether the rates of improvement demonstrated by any of the instructors were significant. For this analysis, we looked at RTOP and Danielson separately because of the difference in scale between the two and to avoid any issues caused by areas where the two measures overlapped. Table IV shows the results of this analysis.

GTA 2 started with the highest scores on both RTOP and Danielson and demonstrated the most improvement over the course of the semester. The other three GTAs began the semester at nearly the same level as one another. GTA 1 showed the least improvement. GTA 4 started with the lowest score and showed the most significant improvement. By both measures, the NTT showed no improvement.

To check whether RTOP and Danielson were measuring similar aspects of teaching, we looked at how well the scores given using RTOP and Danielson correlated with one another. We found they were highly correlated, $R^2 = .831$, $F(1,20) = 98.325$, $p = <.001$, suggesting the constructs measured by each are similar.

B. Comparison to unobserved instructors

In subsequent semesters, we lacked the resources to continue observing our GTAs as often. This provided us with an opportunity to compare the performance of instructors who underwent regular observation and

TABLE IV. Results of regression analysis for RTOP and Danielson observation scores.

RTOP					
Instructor	B	SE(B)	β	t	Sig. (p)
GTA 1	1.948	1.218	0.678	1.599	0.208
GTA 2	6.252	1.554	0.943	4.022	0.057
GTA 3	4.384	1.743	0.825	2.530	0.085
GTA 4	5.439	0.323	0.996	16.829	0.004
NTT	−0.167	0.446	−0.256	−0.374	0.744

Danielson					
Instructor	B	SE(B)	β	t	Sig. (p)
GTA 1	0.701	0.361	0.746	1.942	0.147
GTA 2	1.052	0.127	0.986	8.310	0.014
GTA 3	1.158	0.187	0.963	6.178	0.009
GTA 4	1.117	0.078	0.995	14.386	0.005
NTT	0	0			

instructors who had not. To make this comparison, we observed four GTAs once each at the end of the spring 2022 semester. These observation scores were then compared to the final observation score of each GTA in the Fall 2019 cohort. Figure 5 shows the observation scores from each cohort, with the instructors who were observed and received feedback on the left and instructors who did not on the right.

The average final score of the instructors observed throughout the semester was 0.56 while the average score of the instructors observed only once at the end of the semester was 0.29. This difference was statistically significant ($p = 0.018$). This difference is even more significant considering the scores for the observed instructors are coming from the fall, their first semester teaching, while the scores from the unobserved instructors are coming from the spring, their second semester of instruction, meaning their scores are lower despite having an extra semester of teaching experience.

VIII. QUALITATIVE FINDINGS

In this section, we will flesh out the quantitative findings from the previous sections with qualitative details from our observations and debriefs to characterize the teaching and changes we observed in each instructor throughout the semester.

The ISLE-based labs required several sophisticated skills from the instructors. Because the lab consists of multiple smaller experiments instead of a single experiment taking the entire class time, instructors needed to manage their class time effectively and monitor up to ten student groups at once to judge when the class was ready to proceed from one activity to the next. Because students are expected to develop new ideas instead of simply confirming something they were taught in the lecture, the instructor must facilitate discussions that allow students to share their own ideas rather than giving them a lecture prepared in advance about the physics content of each experiment. Because experiments have different purposes in an ISLE-based class, the instructor must effectively communicate the purpose of each experiment beyond what data students are expected to collect.

Because these lessons were so pedagogically demanding, our observations allowed us to see many different aspects of each instructor's teaching. We would not have learned nearly as much about their teaching ability had they played the more passive role encouraged by traditional

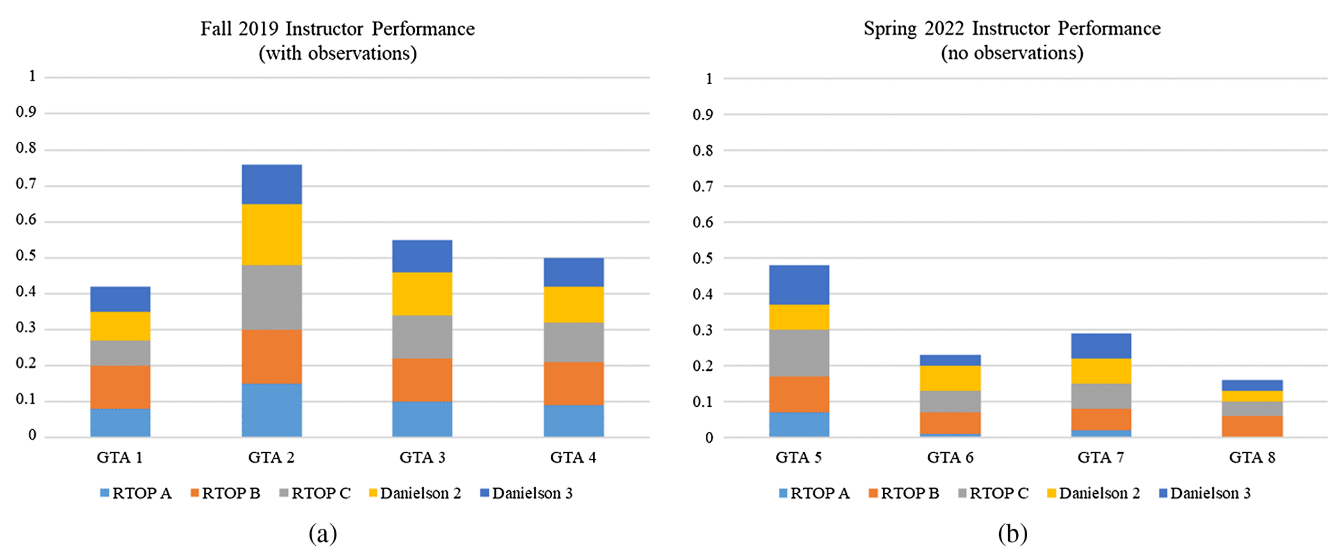


FIG. 5. Comparison of final observation scores of (a) instructors who were observed regularly and (b) instructors who were observed only once.

lab courses. Below, we highlight what we observed from each instructor.

A. GTA 1

Of all the GTAs, GTA 1 was the most confident in his physics content knowledge and, initially, the most skeptical of our teaching methods. A few times during the semester, he challenged certain policies, such as students' being able to revise their work for full credit, or our approach to certain physics topics, such as conceptualizing work entirely as a transfer of energy into or out of a system via some mechanical process. Students also raised several concerns about him privately with members of the reform team, one of which needed to be escalated to a meeting with a supervisor. Despite this, in our interactions with GTA 1 outside of class, he gave the impression that he genuinely wanted to improve his teaching and was open to receiving feedback. He simply required more convincing than the other GTAs that these instructional methods were effective. Below are the themes that emerged throughout our observations of his lessons.

1. Emphasis on grades

GTA 1 often emphasized student grades. As one of the central aspects of the ISLE approach, students were allowed and encouraged to resubmit their work after it had been graded. This means that instructors often spent time at the beginning of a lesson talking about the common mistakes students made on previous assignments and how to fix them. Whenever GTA 1 did this, his focus was on the grades themselves, not the skills or abilities those grades were intended to represent. He told students "If you take [my comments] into account, you can get the highest grade, so this is very, very generous," or "I'm really happy that you all got good grades," or "You all can get the top grade if you resubmit." Often advice he would give students would be phrased as what would get them higher grades rather than what would be correct or appropriate, saying things like "If I don't see a prediction and hypothesis, I can't give you the best score."

2. Soliciting information and asking questions

From the first observation, we noticed a few specific patterns in GTA 1's teaching when it came to asking students questions and soliciting information from them. While he occasionally asked to hear students' reasoning, saying things like "Are the magnitude of these two forces the same? How do you know?" or "Be warned if you put your hand up, I'm going to ask you to explain," in practice, students were given few opportunities to expand on their own ideas during class. During the first observation, GTA 1 would often answer his own questions before giving students a chance to answer for themselves. When students were given opportunities to answer questions, it was often in a call-and-response fashion where the instructor remained the focus of the discussion. Students were never encouraged to respond to one another directly during class discussions.

This was subtly reinforced by the way GTA 1 would ask students to share their findings: "What can you tell me?"

Whiteboard meetings, which were supposed to be a time for groups to share their results with one another, were reduced to each group waiting their turn to report their findings to the instructor. Another way this was reinforced was that GTA 1 would not write down anything students said when answering his questions. This sent the implicit message to students that their ideas were not important or worth recording, regardless of how many times he verbally reminded students to listen to one another.

GTA 1 improved noticeably in this area across the semester. By the third observation, he was employing a technique introduced and modeled in the training meetings of giving students time to confer with the other members of their group and to write down their thoughts on their whiteboards before sharing as opposed to simply asking a question and waiting for students to raise their hands. He also adopted the practice of writing student responses on the board rather than just repeating them verbally or assuming other students heard and memorized them.

3. Pacing and preparation

The pacing of lessons was a common problem for all instructors at the beginning of the semester. For GTA 1, the largest source of pacing problems came from the amount of time he devoted to the whole class discussion at the beginning of each lab. While these discussions were intended to be a short way to introduce the need to know and produce the need to conduct experiments, GTA 1 would regularly begin the first experiment 35 min into his 110-min class period. His discussions often focused on physics content, as they would in a traditional lab where students are expected to verify known physical laws. As a result, after these discussions, students regularly had little context for the experiment they were doing and were confused about the purpose of the experiment. This often led to an inefficient use of time where he would visit each group individually, find out they did not know what they were doing, and give them guidance. With ten groups, this would leave some groups sitting idle for a long time with no way to proceed.

This was an issue that was addressed in training, and GTA 1 did eventually implement the solution we modeled where he would ask students, first expecting a verbal response and later asking them to write on their whiteboards, what they were going to do in an experiment so he could check that they knew their goals before splitting up into their groups to work.

B. GTA 2

GTA 2's teaching was scored consistently higher than that of the other instructors. Perhaps part of the reason she demonstrated so much improvement was that we did not need to spend training time convincing her to adopt student-centered practices and thus had more time to focus on improving specific skills. She was one of the two instructors who took advantage of the optional opportunity to observe the lead instructor's classes.

1. Engaging students in discussion

In her first observed class, GTA 2 wanted to engage students in whole-class discussions but was unable to do so effectively. There were a few choices she made that worked against her. For example, after each experiment, she asked all students to physically join the discussion but did not encourage them to speak to one another. As with GTA 1, the class discussions in her first lesson were very teacher centered. For example, students would hold up their boards and then she would read what was written on them rather than letting the students present their own findings. She insisted that a different member of each group hold the board for each discussion but did not ask them to explain anything for themselves.

Another example of this came from how she tried to ensure that students were listening to one another when they spoke. She recognized that students were unable to hear one another but attempted to solve this problem by summarizing what the students said rather than asking the students to repeat themselves. This undermined her effort to encourage students to pay attention to one another, implicitly teaching students that if another student said something important, the instructor would repeat it for them.

By the second observation, she had made notable improvements. She no longer repeated what students were saying and instead asked them to repeat themselves. However, the whiteboard meetings between experiments were still used predominantly as an opportunity for her to check on students' work rather than as an opportunity for students to present their work to one another. This began changing by the third observation, when she started actively encouraging students to share their boards with the rest of the class, even going so far as to remind students as they prepared that they were meant to be shared and needed to be legible from across the room. She also reminded students during class discussions that they were "talking to each other, not to me."

2. Giving instructions before students were ready for them

In her first observation, the students were using digital force sensors for the first time to take measurements. Because these instruments were new to the students and complex, GTA 2 needed to explain how to use them. This included some complex procedures in the software used to record data such as zeroing the sensors, reversing the direction of one of the sensors, and recording graphs of the collected data. While her explanation was clear, concise, and included a demonstration projected for everyone to see, it was given before students ever had a chance to interact with the probes or software for themselves. Students were shown what to do once and expected to remember everything rather than given a chance to work alongside her. This led to confusion and represented a common theme in her early lessons where students were given information before they were ready to use it.

Something similar occurred in her second observed lesson where she described the shape of a graph that she expected the students to get and talked about how they would analyze their data before the students had even begun collecting data. However, this improved over the course of the semester. By the final observation, she was more comfortable asking students to make small observations first which would inform the discussion she wanted to have rather than rigidly breaking the class time into discussion and experiment segments. She became much more comfortable with a fluid style of instruction which allowed her to more precisely control how her class time was used.

3. Redrawing diagrams

While it is a small point, multiple times throughout the semester GTA 2 either asked students to redraw graphs or diagrams they had already drawn on their whiteboards or redrew them herself. In her first observed lesson, she invited a student up to the large whiteboard at the front of the classroom to redraw the exact same graph the student had already drawn on her group's whiteboard instead of simply using the group's whiteboard. In the final lesson of the semester, GTA 2 did the same thing herself, drawing a force diagram on the board that every group had already drawn on their own whiteboards. This did little except waste class time, which compounded her time management issues. While this issue was pointed out to her each time it was observed, it was not resolved by the end of the semester.

C. GTA 3

GTA 3 showed many of the same patterns as GTAs 1 and 2, most notably having difficulty pacing his lessons. In the first observed lesson, he spent 49 min of his 110-min class in his initial discussion with the students. During that time, the students did not have a chance to do anything but respond to his often very closed questions. By the end of the semester, his initial discussion had been cut down to only 22 min. The fact that he was asking students to sit for such a long time with nothing to do but listen led to discipline problems in his class which he did not have the tools to address. Often, he would pause to tell students to pay attention to him or to listen to what he was saying. The number of discipline problems decreased over the semester as his effective usage of class time improved.

1. Using whiteboards as thinking tools

During his early observations, GTA 3 seemed to view student whiteboards solely as a means of presenting information, both to the instructor and to other groups. Every time he asked students to use their whiteboards, he would follow this by making all of the students form a circle in the middle of the room and present what they wrote. This happened even when asking students to take 3 min to draw a single force diagram. Despite all students drawing the same diagram, he still went through the process of making them present to one another, further compounding his time management problems. This suggests that he did not see the role of whiteboards as facilitating communal thinking and reasoning within groups, but merely as a means of presenting information to others. This contradicted what was taught during the training meetings, where it was repeatedly emphasized that the act of working collaboratively on a whiteboard was valuable even if the resulting work was never presented.

2. Hypothetical examples

A common observation in GTA 3's class was that he would make up hypothetical examples to illustrate the points he wanted to make. While that can be a very effective pedagogical strategy, the examples he chose often confused students more than they helped. Rather than selecting tangible, tactile examples that students could see and feel he would often ask students to imagine the situations he was describing.

In his first observation, he made students imagine a situation where one block was resting on top of another as an example of a situation where friction would cause an object to start moving rather than causing it to stop. Despite students having everything they needed to create this situation for themselves, he only asked them to visualize it. In his final observation, he asked students to imagine pushing on different parts of a door to think about how the distance between where the force is exerted and the axis of rotation affects how hard it is to make the door rotate. He did this not by having any students interact with either of the two doors to the lab room but by drawing diagrams on the whiteboard at the front of the room.

D. GTA 4

GTA 4 was naturally quiet; making himself heard over his students was a frequent struggle throughout the semester. At the start of the semester, this was exacerbated by the fact that his only strategy for dealing with talking students was to attempt to talk over them. While this improved slightly over the course of the semester with GTA 4 learning to wait until he could be heard before he spoke, it was not completely resolved by the end of the semester. GTA 4 was the second instructor who took the opportunity to observe the lead instructor's classes on multiple occasions.

1. Pacing and time management

Like the other instructors, GTA 4 struggled to maintain an appropriate pace in his classes at the beginning of the semester. However, unlike GTAs 1 and 3 who spent an excessive amount of time lecturing students at the beginning of each class, GTA 4 instead tried to actively engage his students during that time. He regularly included group whiteboard activities at the beginning of each lesson to get students warmed up, but his sense of timing for these activities was off at the beginning of the semester. He was hesitant to cut any student off while they were still working which resulted in these whiteboarding activities taking significantly longer than they needed to. He had the same reluctance to end experiments, leading those to drag on as well. However, after concrete methods for dealing with this such as giving students explicit time limits before each activity was included in the training meetings, he began implementing them in his class and his time management improved.

2. Ensuring students understood their goals

As with the other instructors, in his earliest observations, GTA 4 sent students to conduct their experiments without ensuring they understood what they were trying to do or why. He began the first experiment of his first observed lesson by telling students "If you know everything, please start doing your experiment" without checking to see whether they understood what they were doing.

After the importance of ensuring students understood their goals was emphasized in training and strategies for checking for this understanding were discussed, GTA 4 implemented those methods in his class. He would ask students to write their goals on their whiteboards before beginning their experiment or he would solicit information from multiple students rather than being satisfied when the first student he called on gave a positive response.

However, the goals that he wanted students to understand were often superficial, limited to what students were doing rather than why they were doing it. In one experiment, students were testing a hypothesis about the conditions necessary for static equilibrium by predicting where along its length a meter stick with a block hanging from one end would need to be supported in order to keep it from falling or rotating. While he ensured that all students knew they were supposed to predict the location of the support, following up with his students revealed that none of them could explain why they were making that prediction. While getting students to identify what they were trying to do was a step in the right direction, there was still room for improvement by the end of the semester.

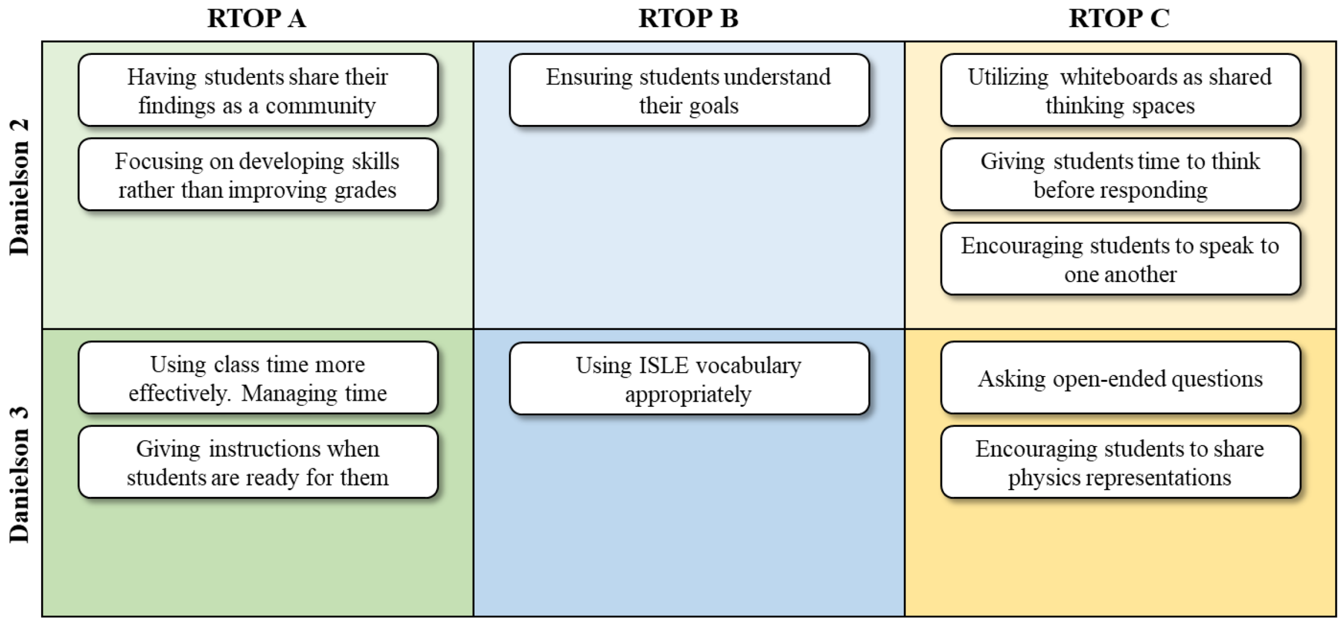


FIG. 6. Overview of classroom practices in which GTAs demonstrated improvement, categorized by the quantitative domains with which they align.

E. NTT

Many of the same issues observed in the GTAs' classes were observed with the NTT, but unlike with the GTAs, few changes were observed in the NTT's instruction across the semester. He did not provide students with any context or means of understanding why they were doing each experiment. To deal with the issue of managing time during his lessons, the NTT decided to leave out the whiteboard meetings in his classes. Instead, he looked over each group's work individually before allowing them to proceed to the next experiment independently. His reasoning for this was that students finished each experiment at different times, and he did not see the value in the whiteboard meetings themselves. As a result, there was no sense of community in his classes. Groups of students would leave class at different times whenever they finished the final experiment, an occurrence common in traditional labs but not observed in any of the other instructors' classes.

1. Rejecting feedback

After each observation, attempts were made to give the NTT feedback on his lesson. However, this feedback was shrugged off every time. The NTT would explain why particular activities or ways of teaching, such as the whiteboard meetings, were impossible to manage. If he did not think that something was impossible, then he insisted it had no value and did not want to include it in his teaching.

F. Summary of qualitative findings

While each instructor improved at their own rate and in different areas, there was overlap in the improvements we observed. Figure 6 highlights some of the classroom practices in which many of our TAs demonstrated observable improvement over the course of this study. We have situated each practice within the domain used for the qualitative analysis with which it is most closely aligned.

IX. DISCUSSION

We began this study with three questions. In this section, we will discuss how our data answers these questions.

A. What aspects of teaching were most difficult for our instructors?

Having never taught classes in any official capacity before, our instructors struggled with many of the basic tasks of classroom teaching. At the beginning of the semester, they had difficulty with time management and facilitating group discussions. They struggled to ask open-ended questions which invited students to share their thoughts with the class and to hold students accountable for listening to each other.

In terms of the ISLE approach, there were three key areas where our instructors seemed to struggle: using ISLE-based vocabulary correctly, fostering a community of learners, and making students aware of the epistemic purpose of each activity.

1. Confusion about ISLE-based vocabulary

The vocabulary used in the ISLE approach is intended to make the inductive and hypothetico-deductive processes of scientists transparent for students. When these terms are used correctly, students do not need to wonder why they have hypotheses in some experiments and not others or why they only make predictions in certain experiments. Similarly, the ISLE approach uses the terms hypothesis and prediction—terms that often get muddled together or used interchangeably in other formulations of scientific inquiry—in very specific and distinct ways. The correct and consistent use of this vocabulary constitutes a critical component of ISLE-based instruction. Often throughout the semester, especially at the beginning of the year, we saw instructors confusing predictions with hypotheses or treating testing experiments as if they were observational experiments.

2. Building a community of learners

While this is not unique to the ISLE approach, the understanding that physics is a sociocultural process [62] serves as one of the main tenets underlying the ISLE approach. One of the purposes of the whiteboard meetings between each experiment is to foster this understanding among students. In professional science, findings must be communicated, disagreements negotiated, and consensus reached. Establishing this kind of classroom culture is exceedingly difficult, even for experienced instructors. It requires not just pedagogical skill, but trust and a willingness to give students ownership over the learning process.

All our GTAs were willing to attempt to develop this classroom culture from the very beginning of the year, though they had varying levels of success. They all attempted to hold whole-class discussions between experiments while the NTT was not. Even when they felt like their time management problems pressed them, they did not abandon this aspect of the course. Though they were willing to try, many of their attempts began in a teacher-centered way, with them essentially asking each group to report their findings directly to the instructor and little communication between groups. This is not an unexpected place for new instructors to begin, but it is clear they have much room for improvement.

3. Making students aware of the epistemic purposes of activities

In the ISLE approach, "every activity has an identifiable epistemological (knowledge-generating) purpose." [61] (p. 5) When our curricular materials were developed, each experiment was meant as part of an intentional process where students develop, test, refine, and apply their own ideas to construct new physics knowledge. However, just because this was the curriculum developer's intent does not mean it translated into the classroom practices of the instructors. Whether because our instructors did not understand the intended purposes of the activities themselves or because they did not have the skills to make students aware of those purposes, many of the lessons we observed included activities whose epistemic purposes were unknown to students.

As mentioned earlier, all of the new instructors started the semester by asking students to conduct experiments whose purposes they did not make clear. We have already discussed the effect this had on their time management, but it also undercuts the effectiveness of the ISLE approach. One of the problems commonly cited about traditional lab activities is the passivity and lack of agency of students; as Hoffstein and Lunesta described, "many students engage in laboratory activities in which they follow recipes and gather and record data without a clear sense of the purpose and procedures of their investigation and their interconnections." [63] (p. 40) While the design of our curriculum greatly reduced the number of "recipes" available for students to follow, even when students were tasked with designing their own experimental procedures, they were given only superficial goals for the procedures they were asked to design. For example, the students were told to design a procedure that would allow them to collect specific data without understanding why collecting those data was important.

B. How did the teaching practices of our instructors change over the semester?

While some showed more growth than others, all of our GTAs became more proficient teachers over the course of the semester. In particular, we saw all GTAs begin to adopt specific, concrete practices shown to them during our training meetings such as announcing that students had specific time limits for each activity, using the group white boards as a collaborative brainstorming space before asking students questions, and asking students to repeat the goals for each activity before allowing them to start.

As a result, all four of our GTAs were more proficient teachers at the end of the semester than they were at the beginning. Their time management improved, and while they still struggled to facilitate group discussions where students talked to one another rather than just talking to the instructor, there were noticeable improvements in the kinds of questions they asked students as well as how these questions were asked.

Their understanding of ISLE-specific vocabulary and the ISLE process improved as well. We observed fewer instances of instructors confusing hypotheses with predictions as the semester went on, and instructors seemed to better understand the difference between observational experiments and testing experiments. We did not expect any of our instructors to become experts in the ISLE approach after the single semester of this study. When training preservice physics teachers in the ISLE approach, it takes repeated exposure to these ideas and vocabulary over the course of a 2-year master's program [64] for teachers to approach using these ideas naturally. It would be unreasonable to expect our new instructors to completely internalize this new way of thinking and speaking about experimentation after a single presemester training session and one semester of weekly training meetings.

Another significant improvement we observed was the GTAs becoming more conscientious about making sure their students knew the goals of each activity before they started working. Over time, we saw all of the GTAs making more active efforts to ensure that students understood what they were supposed to do, which was an improvement from the beginning of the semester, but they often fell short of ensuring students understood why they were doing each experiment. Recall the episode from GTA 4's lab on static equilibrium; he made sure that students understood they were supposed to make a prediction but did not discuss why students were being asked to make that prediction. So, while our instructors learned that they were supposed to convey the goals of each activity to their students, they were not always able to distinguish between the practical goals (what they wanted students to be doing) and epistemic goals (why the activity was being done) of the lab activities.

C. Which training practices had the largest effect on our instructor's teaching practices?

We certainly do not attribute all of our GTAs' growth as instructors to the training we provided. New instructors naturally improve as they adjust to their roles. However, the specific patterns of improvement we observed suggest that some of our training practices were particularly effective at changing our GTA's teaching practices.

Our training meetings consisted of a combination of theoretical pedagogical discussions, working through the activities they were going to teach, and sharing concrete practices the instructors could use in their classes. Of these, the concrete practices had the most immediate impact on how our instructors performed, being adopted by all GTAs as soon as they were introduced. This is not to suggest that the sharing of these classroom tips and tricks is more important than other elements of professional development, or that training consisting of only sharing such techniques would be effective, only this was the element of our training that had the most visible and immediate effect. We speculate that there were two main reasons for this: these practices were easy to implement, and they addressed the instructors' immediate concerns. Asking questions in a different way or giving students time to confer among themselves before soliciting an answer to your question does not require you to rethink your approach to teaching, nor does giving students an explicit time limit for each activity. By design, these techniques were meant to be easy to implement. Furthermore, they were inspired by our observations. We did not plan to include any specific techniques in our training meetings. Instead, we introduced them as we saw the need for them. When we observed that many of our instructors were struggling with time management, we shared small tricks to help with that. The fact that these techniques were a response to problems that the instructors were having in their classes provided them with motivation to adopt them.

This leads to the importance of the observations themselves. While originally the observations were only intended for research purposes, we found that performing these observations significantly improved the effectiveness of our training compared to the other semester when we observed our instructors only once. We believe there are two primary reasons for this. First, the observations and subsequent debriefings allow for more personalized feedback where instructors are told what they did well and what they should focus on. The observers were able to bring things to the instructors' attention that they had not noticed while teaching. Second, the act of observing all of the instructors allowed us to identify areas where many of our instructors were struggling and customize our training accordingly. Since we were in their classrooms, we knew what difficulties the GTAs were facing and could plan our training meetings accordingly.

D. Why did the NTT not change?

While all our GTAs improved over the course of this study, the NTT's teaching did not change at all. We felt that it was important to include him in this study as a counterexample to show that improvement is not guaranteed. While we do not know for sure why he did not change, the most notable thing we observed was his resistance to feedback. While all of the GTAs were willing to try using the ISLE approach, the NTT preferred to run his class as if it were a traditional laboratory course, conducting all of the same experiments but leaving out all aspects of the course meant to build community and engage students in the process of developing scientific ideas. When challenged he would either say that he did not see the value of those aspects or claim they could not be done. This is consistent with prior research emphasizing the importance of an instructor's beliefs about education matching those inherent to the style of instruction [65,66].

X. IMPLICATIONS FOR INSTRUCTOR TRAINING

Our findings suggest that observations and mentoring new instructors play an important role in their early professional development. These observations provide opportunities for new instructors to reflect on situations and problems that arise in their class with an experienced educator and can be used to inform the formal training they receive. Such practice allows the training to cater to the immediate needs of the instructors, making it seem more relevant to them. Many of the specific strategies GTAs adopted were added to the training meetings specifically to address problems we saw during our observations.

While more general philosophical and pedagogical knowledge is important and should be included in any instructor training, we found that concrete strategies instructors could implement in their classes immediately produced the biggest changes in their performance, such as giving students concrete time limits for each activity, giving students time to confer with their neighbors and on whiteboards before answering questions, and recording students' verbal responses on the whiteboard at the front of the room during class discussions. The fact that these strategies were selected because they addressed problems that the instructors were already facing in their classes likely increased their adoption.

Finally, attention should be paid to the beliefs new instructors have about education and what their role in the classroom should be. If instructors have beliefs that are incompatible with the teaching methods you are presenting, it is likely they will resist or refuse to change. We were fortunate that all our GTAs were willing to accept the ISLE approach, but our NTT provides an instructive example of what happens when this is not the case.

A. How to support novice instructors teaching an ISLE-based lab course

We believe that many of the recommendations we can make based on our results are generally good practices for training any novice instructors, not simply for those tasked with teaching an ISLE-based course. However, here we will outline what we believe were the most impactful elements of our training, those we would recommend anyone looking to train novice instructors to teach student-centered classes implement.

Weekly training meetings where instructors can work through class activities as students are essential. The activities students complete in an ISLE-based course are unlike those an instructor with a background in traditional classes saw when they were a student. During these meetings, it is important to constantly reinforce ISLE-specific vocabulary so instructors will pick it up and use it correctly with their students. The differences between an observational experiment, as well as between a hypothesis and a prediction, take time even for instructors to internalize. During these meetings, the lead instructor should model all the behaviors instructors will be expected to use in their classes while being explicit about what they are doing and why it is important. Do not expect novice instructors to pick up on the specifics of your methods on their own simply by watching you teach.

Periodic observations of new instructors by the lead instructor or another qualified faculty member are essential for the quick development of new instructors and serve many roles. They allow novice instructors to get real-time feedback and guidance when problems arise, they allow the lead instructor to tailor future training to meet the needs of novice instructors more directly, and they provide opportunities for novice instructors to receive detailed feedback and to reflect on their teaching. If it is feasible, co-teaching with a more experienced instructor would accomplish these same goals.

Finally, just as when working with students, it is important that we be transparent about our goals and rationales when communicating with novice instructors. Explaining why we are asking them to teach the way we do, and how it affects students and facilitates learning, can go a long way toward getting novice instructors to buy into methods that may seem foreign and unnecessarily complicated to them coming from a traditional background.

XI. LIMITATIONS AND AREAS OF FUTURE RESEARCH

This study involved a narrow sample of five instructors, four of whom were completely new to teaching, all of whom were teaching the same course. The lab curriculum they were responsible for teaching was developed by the same researcher who observed their classes and led their weekly training meetings. This, in some ways, presents an ideal situation for training. While it is possible that lessons learned from this study may be applicable to broader contexts of new instructor training, this applicability has not been conclusively demonstrated by this study. The observer and trainer was a member of the instructional team teaching the same course. This may limit the applicability of these results when the observations or training come from administrators or outside professional development programs.

Another limitation of this study came from the participants not recording their experiences. Because of this, many of our claims about why certain practices were adopted are only speculation. It is strongly recommended that journaling and reflection be made an institutionalized part of new instructors' responsibilities to avoid this in the future. This is not limited to the study participants. The researchers and participants worked alongside one another as part of the same teaching team. There were many informal interactions between the researchers and instructors which went undocumented. While any interactions deemed to be important were documented shortly after, this was not possible for every interaction. We cannot discount the possibility that some of these interactions had impacts on the participants which went unrecorded.

[1] C. Henderson, Promoting instructional change in new faculty: An evaluation of the physics and astronomy new faculty workshop, Am. J. Phys. **76**, 179 (2008).

[2] W. Sun and S. Simon-Roberts, New faculty preparation, adaptation, and retention, J. Fac. Dev. **34**, 81 (2020).

[3] P. Knight, *Being a Teacher in Higher Education* (McGraw Hill Education, Buckingham, 2002).

[4] E. Etkina, Millikan award lecture: Students of physics—Listeners, observers, or collaborative participants in physics scientific practices?, Am. J. Phys. **83**, 669 (2015).

[5] E. Etikina, D. T. Brooks, and G. Planinsic, *Investigative Science Learning Environment: When Learning Physics Mirrors Doing Physics* (Morgan & Claypool Publishers, San Rafael, 2019).

[6] 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).

[7] L. B. Wheeler, C. P. Clark, and C. M. Grisham, Transforming a traditional laboratory to an inquiry-based course: Importance of training TAs when redesigning a curriculum, J. Chem. Educ. **94**, 1019 (2017).

[8] W. S. Toothacker, A critical look at introductory laboratory instruction, Am. J. Phys. **51**, 516 (1983).

[9] K. Tobin, Research on science laboratory activities: In pursuit of better questions, and answers to improve learning, School Sci. Math. **90**, 403 (1990).

[10] R. Trumper, The physics laboratory—A historical overview and future perspectives, Sci. Educ. **12**, 645 (2003).

[11] N.-H. Kang and C. S. Wallace, Secondary science teachers' use of laboratory activities: Linking epistemological beliefs, goals, and practices, Sci. Educ. **89**, 140 (2006).

[12] C. E. Wieman and N. G. Holmes, Measuring the impact of an instructional laboratory on the learning of introductory physics, Am. J. Phys. **83**, 972 (2015).

[13] B. R. Wilcox and H. J. Lewandowski, Open-ended versus guided laboratory activities: Impact on students' beliefs about experimental physics, Phys. Rev. Phys. Educ. Res. **12**, 020132 (2018).

[14] American Association of Physics Teachers (AAPT), *Recommendations for the Undergraduate Physics Laboratory Curriculum* (American Association of Physics Teachers, 2014).

[15] N. G. Holmes, C. E. Wieman, and D. A. Bonn, Teaching critical thinking, Psychol. Cognit. Sci. **112**, 11199 (2015).

[16] J. Jackson, Modeling instruction: An effective model for science education, Sci. Educ. **17**, 10 (2008), https://files.eric.ed.gov/fulltext/EJ851867.pdf.

[17] R. J. Beichner and J. M. Saul, Introduction to the SCALE-UP (Student-Centered Activities for Large Enrollment Undergraduate Programs) project, in *Proceedings of the International School of Physics "Enrico Fermi"*, (Venice, Italy, 2003), https://www.researchgate.net/profile/Robert-Beichner/publication/242298460_Introduction_to_the_SCALE-UP_Student-Centered_Activities_for_Large_Enrollment_Undergraduate_Programs_Project/links/00b7d53c56fc50a75c000000/Introduction-to-the-SCALE-UP-Student-Centered-Activities-for-Large-Enrollment-Undergraduate-Programs-Project.pdf.

[18] N. G. Holmes and H. J. Lewandowski, Investigating the landscape of physics laboratory instruction across North America, Phys. Rev. Phys. Educ. Res. **16**, 020162 (2020).

[19] M. Riaz, T. J. Marcinkowsi, and A. Faisal, The effects of a DLSCL approach on students conceptual, Eurasia J. Math. Sci. Technol. Educ. **16**, em1813 (2020).

[20] R. L. Kung, Teaching the concepts of measurement: An example of a concept-based laboratory course, Am. J. Phys. **73**, 771 (2005).

[21] A. E. Lawson, C. B. McElrath, M. S. Burton, B. D. James, R. P. Doyle, S. L. Woodward, L. Kellerman, and J. D. Snyder, Hypothetico-deductive reasoning skill and concept acquisition: Testing a constructivist hypothesis, J. Res. Sci. Teach. **28**, 953 (1991).

[22] E. Etkina, A. Van Heuvelen, S. White-Brahmia, D. T. Brookes, M. Gentile, S. Murthy, D. Rosengrant, and A. Warren, Scientific abilities and their assessment, Phys. Rev. ST Phys. Educ. Res. **2**, 020103 (2006).

[23] A. E. Lawson, The generality of hypothetico-deductive reasoning: Making scientific thinking explicit, Am. Biol. Teach. **62**, 482 (2000).

[24] C. S. Dweck and D. S. Yeager, What can be learned from growth mindset controversies?, Am. Psychol. **75**, 1269 (2020).

[25] E. Wenger, Communities of practice and social learning systems, Organization **7**, 225 (2000).

[26] D. Rosengrant, E. Etkina, and A. Van Heuvelen, An overview of recent research on multiple representations, AIP Conf. Proc. **883**, 149 (2007).

[27] R. Boice, New faculty as teachers, J. Higher Educ. **62**, 150 (1991).

[28] R. E. Cipriano and J. L. Buller, Rating faculty collegiality, Change **44**, 45 (2012).

[29] M. H. Dancy and C. Henderson, Experiences of new faculty implementing research-based instructional strategies, AIP Conf. Proc. **1413**, 163 (2012).

[30] R. H. Stupinsky, M. B. Weaver-Hightower, and Y. Kartoshkina, Exploring and testing the predictors of new faculty success: A mixed methods study, Stud. Higher Educ. **40**, 368 (2015).

[31] S. Ambrose, T. Huston, and M. Norman, A qualitative method for assessing faculty satisfaction, Res. High. Educ. **46**, 803 (2005).

[32] M. Murtonen and H. Vilppu, Change in university pedagogical culture—The impact of increased pedagogical training on first teaching experiences, Int. J. Learn. Teach. Educ. Res. **19**, 367 (2020).

[33] T. L. Derting, D. Ebert-May, T. P. Henkel, J. Middlemis Maher, B. Arnold, and H. A. Passmore, Assessing faculty professional development in STEM higher education: Sustainability of outcomes, Sci. Adv. **2**, 1 (2016).

[34] D. Ebert-May, T. L. Derting, T. P. Henkel, J. Middlemis Maher, J. L. Momsen, B. Arnold, and H. A. Passmore, Breaking the cycle: Future faculty begin teaching with learner-centered strategies after professional development, CBE Life Sci. Educ. **14**, ar22 (2017).

[35] C. Henderson and M. H. Dancy, Barriers to the use of research-based instructional strategies: The influence of

both individual and situational characteristics, Phys. Rev. ST Phys. Educ. Res. **3**, 020102 (2007).
[36] R. Hake, Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, Am. J. Phys. **66**, 64 (1997).
[37] C. E. Wieman, Large-scale comparison of science teaching methods sends clear message, Proc. Natl. Acad. Sci. U.S.A. **111**, 8319 (2014).
[38] L. S. Shulman, Those who understand: Knowledge growth in teaching, Educ. Res. **15**, 4 (1986).
[39] D. L. Ball, M. H. Thames, and G. Phelps, Content knowledge for teaching: What makes it special?, J. Teach. Educ. **59**, 389 (2008).
[40] J. E. Zull, *The Art of Changing the Brain: Enriching Teaching by Exploring the Biology of Learning* (Stylus Publishing, Sterling, VA, 2002).
[41] C. T. Fosnot and R. S. Perry, Constructivism: A psychological theory of learning, in *Constructivism: Theory, Perspectives, and Practice* (Teachers College Press, New York, NY, 1996), Vol. 2, pp. 8–33.
[42] S. Magnusson, J. Krajcik, and H. Borko, Nature, sources, and development of pedagogical content knowledge for science teaching, in *Examining Pedagogical Content Knowledge* (Springer, Netherlands, 1999), pp. 95–132.
[43] A. D. Robertson, Supporting the development of curricular knowledge among novice physics instructors, Am. J. Phys. **86**, 305 (2018).
[44] L. C. McDermott, P. S. Shaffer (The Physics Education Group at University of Washington), *Tutorials in Introductory Physics* (Prentice Hall, Upper Saddle River, 1998).
[45] N. R. Council, *A Framework for K-12 Science Education: Practices* (The National Academy Press, Washington, DC, 2012).
[46] E. Etkina, Pedagogical content knowledge and preparation of high school physics teachers, Phys. Rev. ST Phys. Educ. Res. **6**, 020110 (2010).
[47] E. Etkina, Using early teaching experiences, and a professional community to create habits of student-centered instruction, and to prevent attrition, in *Recruiting, and Educating Future Physics Teachers*, edited by C. Sandifer and E. Brewe (American Physical Society, College Park, MD, 2015), pp. 249–266.
[48] N. G. Holmes, M. Martinuk, J. Ives, and M. Warren, Teaching assistant professional development by and for TAs, Phys. Teach. **51**, 218 (2013).
[49] E. Alicea-Muñoz, C. S. Sullivan, and M. F. Schatz, Transforming the preparation of physics graduate teaching assistants: Curriculum development, Phys. Rev. Phys. Educ. Res. **17**, 020125 (2021).
[50] T. C. Pentecoast, L. S. 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), https://www.nsta.org/resources/research-and-teaching-graduate-teaching-assistant-training-fosters-student-centered.
[51] L. B. Wheeler, Transforming a traditional laboratory to an inquiry-based course: Importance of training TAs when redesigning a curriculum, J. Chem. Educ. **94**, 1019 (2017).
[52] S. Kanim and X. Cid, Demographics of physics education research, Phys. Rev. Phys. Educ. Res. **16**, 020106 (2020).
[53] E. Etkina, S. Murthy, and X. Zou, Using introductory labs to engage students in experimental design, Am. J. Phys. **74**, 979 (2006).
[54] E. Etkina, A. Van Heuvelen, D. T. Brooks, and D. M. Mills, Role of experiments in physics instruction—A process approach, Phys. Teach. **40**, 351 (2002).
[55] C. Megawan-Romanowicz, Whiteboarding: A tool for moving classroom discourse from answer-making to sense-making, Phys. Teach. **54**, 83 (2016).
[56] B. L. Smith, Taking structure seriously: The learning community model, Lib. Educ. **77**, 42 (1991).
[57] D. Sawada, M. D. Piburn, E. Judson, J. Turley, K. Falconer, R. Benford, and I. Bloom, Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol, School Sci. Math. **102**, 245 (2002).
[58] C. Danielson, *Enhancing Professional Practice: A Framework for Teaching* (Association for Supervision and Curriculum Development, Alexandria, VA, 2007), 2nd ed.
[59] C. Wieman and S. Gilbert, The Teaching Practices Inventory: A new tool for characterizing college and university teaching in mathematics and science, CBE Life Sci. Educ. **13**, 552 (2014).
[60] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.19.020153 for observation protocol used to assess instruction in this study.
[61] J. W. Creswell and V. L. P. Clark, *Designing and Conducting Mixed Methods Research* (Sage Publications Inc., Thousand Oaks, CA, 2007).
[62] D. T. Brooks, E. Etkina, and G. Planinsic, Implementing an epistemologically authentic approach to student-centered inquiry learning, Phys. Rev. Phys. Educ. Res. **16**, 020148 (2020).
[63] A. Hoffstein and V. N. Lunesta, The laboratory in science education: Foundations for the twenty-first century, Sci. Educ. **88**, 28 (2004).
[64] E. Etkina, Pedagogical content knowledge and preparation of high school physics teachers, Phys. Rev. ST Phys. Educ. Res. **6**, 020110 (2010).
[65] N.-H. Kang and C. S. Wallace, Secondary science teachers' use of laboratory activities: Linking epistemological beliefs, goals, and practices, Sci. Educ. **89**, 140 (2005).
[66] A. D. Roehrig, J. E. Turner, C. G. Grove, N. Schneider, and Z. Liu, Degree of alignment between beginning teachers' practices and beliefs about effective classroom practices, Teach. Teach. Educ. **44**, 164 (2009).