Lab instruction during the COVID-19 pandemic: Effects on student views about experimental physics in comparison with previous years

Michael F. J. Fox,* Jessica R. Hoehn, Alexandra Werth, and H. J. Lewandowski

JILA, National Institute of Standards and Technology and University of Colorado, Boulder, Colorado 80309, USA

and Department of Physics, University of Colorado, Boulder, Colorado 80309, USA

(Received 9 March 2021; accepted 3 June 2021; published 30 June 2021)

Physics lab instructors were forced to adapt their courses in 2020 due to the COVID-19 pandemic. We investigate the impact these changes had on student views towards experimental physics as measured by the Colorado Learning Attitudes about Science Survey for Experimental Physics (E-CLASS). Analysis of the responses from over 1600 students in both spring and fall semesters and performing a comparison with the same courses in 2019 shows that student total E-CLASS scores were not lower in 2020 compared to 2019. Nevertheless, in the Fall 2020 data, we find that there is a variation in the mean E-CLASS scores on some individual questions when compared to previous years.

DOI: 10.1103/PhysRevPhysEducRes.17.010148

I. INTRODUCTION

The coronavirus pandemic impacted all aspects of life in 2020. With many educational institutions transitioning to remote learning to reduce the transmission of the virus, undergraduate student access to physics teaching laboratories was severely curtailed. The dramatic disruption during 2020 is expected to have a large, negative impact on student learning [1,2]. While there is some early evidence documenting such an impact, it has focused on student learning of content at the K-12 level [3–5]. In contrast, in this work, we focus on students in higher education, specifically those taking first-year (or introductory level) courses in experimental, laboratory physics. It is the aim of this work to identify the impact of disruption in lab instruction due to the pandemic on student epistemologies and expectations of experimental physics; together, we refer to these as "views" towards experimental physics. Furthermore, we are specifically interested in what impact can be measured at the large scale, identifying changes that are shared across multiple U.S. universities and colleges.

The study of physics labs in the context of the pandemic is particularly important, because these are classes where group work and hands-on experience feature prominently [6,7], and, almost by definition, rely on the class being held in person. Indeed, in a previous report [8], we presented preliminary results indicating that maintaining or replicating the in-person experience was a key challenge when instructors were forced to transition to emergency remote teaching in March 2020. Adapting to this challenge has been a monumental undertaking by instructors, often supported by their peers in the wider community of lab instructors and researchers who shared their experiences and knowledge of best practice throughout 2020 [9–13].

We are interested in how students' views have been affected by emergency remote teaching, rather than student understanding of physics concepts or content knowledge, which may differ widely across different courses and educational settings. A common goal for undergraduate physics lab classes is the development of the habits of mind of an experimental physicist [14,15], which are directly related to student views about experimental physics. Student views are shaped over their entire lifetime of experiences inside and outside the classroom [16], therefore, as we only consider the impact of a single course, we would normally not expect to see large changes in our measure of student views. However, the scale of the changes implemented in 2020 (practically all instructors across the U.S. were changing their courses at the same time) in combination with both students and instructors dealing with the physical and emotional impact of a pandemic, might result in larger changes.

We choose to approach this investigation by adopting a situated theory of epistemological development [17], whereby students' epistemologies develop based on reflections of their experiences in specific cultural and social contexts. Applying this idea to lab courses suggests that the absence of the context of the in-person lab will lead to less relevant experiences for students to reflect upon and, therefore, fewer opportunities for epistemological

*michael.fox@colorado.edu

development. Hence, we propose the hypothesis that student views toward experimental physics will become less expertlike during emergency remote instruction to a greater extent than during in-person, pre-pandemic instruction. We will test this hypothesis with student responses to the Colorado Learning Attitudes about Science Survey for Experimental Physics (E-CLASS) [14].

Since 2012, the E-CLASS has been used by lab instructors to measure how their course affects the development of students' views by administering a survey at the beginning (the pretest) and end (the post-test) of the course (see Sec. II A for more details). While many regular data collection efforts, such as standardized testing and in-person research have been hindered by the public health restrictions imposed by the pandemic [18], as the E-CLASS is administered using an automated, online system [19], we have been able to collect data throughout 2020. In both the spring and fall semesters of 2020, the E-CLASS was used in over 50 courses with responses each semester from over 2500 students. We choose to analyze and discuss the two semesters separately, as there were material differences between how courses were implemented between the two. In spring, courses switched, sometimes in a matter of days, from being in-person to remote teaching, while in fall, many courses were entirely remote or hybrid and instructors had a slightly longer period of time in order to plan how the activities would be realized.

In this work, we focus on the broad question of whether there was an impact on student views. Specifically, we pose the following research questions that will allow us to test our hypothesis:

RQ1. Did students become more or less expertlike in their views about experimental physics in 2020 compared to previous years?

RQ2. What were the students' views about experimental physics that changed in 2020 compared to previous years?

As well as providing documentary evidence as to the impact of the pandemic on students' views around experimental physics, the answers to these research questions will allow the community to evaluate the effectiveness of emergency remote instruction and whether it is worth considering if any aspects of the remote activities should be kept in future courses. However, we save the question of which specific instructional changes during the pandemic were the most successful for later work, where qualitative data sources will be used to build a more complete picture of emergency remote instruction.

The structure of this work is as follows. In Sec. II, we describe first the data collection methods using the ECLASS; how we have selected appropriate data for the comparisons needed to answer our research questions; we report the resulting demographic information of the selected sample of students, then discuss the limitations of the methodology and the ethical considerations of conducting research during a pandemic. In Sec. III, we present our results, first for the spring semester and then for the fall semester. For each semester, we present data on overall E-CLASS scores to answer RQ1 and then present data for individual questions from the E-CLASS to answer RQ2. In Sec. IV, we discuss the results, draw conclusions in answer to our research questions and evaluate the degree to which the evidence presented supports our hypothesis. See Supplemental Material [20] for Appendices A–C.

II. METHODOLOGY

A. The E-CLASS

We start by briefly summarizing the structure, logistics, and processing of the E-CLASS survey. More details of these can be found in Refs. [14,19]. The E-CLASS asks students the same 30 Likert-style questions at the beginning (the pretest) and end (the post-test) of their course. These 30 questions have two parts, the first asking about the student's own experience in their lab course, while the second part asks about how they think expert physicists would respond to the same prompt. We are considering only the former subset of questions in this study, as we are interested only in their views about experimental physics related to *their course*. An additional set of questions are included in the post-test asking students about grading practices in their course, as well as to collect demographic information. During 2020, we also included supplementary questions in the post-test to learn about student experiences of labs during the pandemic, examples of which can be found in Ref. [8]. A simple analysis of a subset of the supplementary remote lab survey questions has been performed in Appendix A [20], the purpose of which is to provide the reader with the necessary context of the courses we are using in the analysis. These descriptions are presented in Sec. III, as a prelude to our presentation of the quantitative results. We choose not to look for associations between this simple analysis of contexts and the E-CLASS scores in this work, as this is beyond the scope of our research questions.

Each of the 30 questions in the E-CLASS that ask about the student's own experience provide a statement to the student which they are asked whether they strongly disagree, disagree, neutral, agree, or strongly agree with. For scoring, this five-point scale is reduced to a three-point scale by collapsing the two extremes at each end of the scale. Then, the student's response is compared with the expertlike response (determined by expert consensus [14]) and given a score of 1 if the two are aligned (i.e., both agree or both disagree), -1 if the two are opposite (one agrees and the other disagrees), and 0 if the student responded with the neutral response. The total E-CLASS score is then calculated by the sum of the scores given for each of the 30 items, therefore existing as integers in the range $[-30, 30]$.

As we are interested in the impact of instruction, it is important to be able to compare a student's score from the

post-test with the score from the pretest. Therefore, we match student responses from the two surveys using their student identification number or their first and last names, which they are asked to provide at the beginning of each survey.

Before proceeding, we highlight the results of one previous study looking at E-CLASS scores to help establish reader expectations for the values presented later in Sec. III, and also to foreground some of our discussion in Sec. IV. As part of the study, Wilcox and Lewandowski [21] compared pretest and post-test scores from the E-CLASS collected from 49 different first-year courses. They reported that in traditional, guided, first-year lab courses there was a drop of 1.9 points in the mean total E-CLASS score from the pretest to the post-test. This drop corresponded to a small effect size of $r = 0.1$ (we discuss our measure of significance in Sec. II B immediately following). Such drops in scores are common on attitudinal surveys in both traditional lab and lecture courses [22]. For first-year lab courses with open-ended activities the mean total E-CLASS score was practically the same from pretest to post-test (it increased by 0.1 points and was not a statistically significant difference).

B. Data identification and analysis

The research questions posed in Sec. I require us to make a comparison between scores on the E-CLASS in 2020 with scores from previous years. To control, to as great an extent as possible, for variations between courses, we include courses only if we have data collected from them when they ran in both 2019 and 2020. We also treat the spring and fall semesters as separate sets of data because of the distinct situations in each semester in 2020, as discussed in Sec. I. Additionally, we have selected data from students in the same courses that ran in 2018, with the purpose of evaluating the size of changes one might expect in a comparison between two years (2018 and 2019) in the absence of the structural changes made because of the pandemic. However, as we do not have data for all the 2020 courses in 2018, this results in fewer courses available to analyze from 2018. This will have an impact on the validity of the comparison, though it remains useful, as it can be used to construct an upper limit on the size of the effects we would expect to see. Hence, our main conclusions are drawn from comparisons between 2020 and 2019 only.

We have chosen to analyze only first-year (or introductory-level) courses, as these form the dominant component of the data in terms of numbers of students and courses. Previous work has shown significant differences in scores on the E-CLASS between first-year and beyond-first-year courses [23,24]. As beyond-first-year courses tend to have fewer students than first-year courses, our ability to draw conclusions from the small dataset of beyond-first-year scores that we have (78 total students in Spring 2020 and 7 total students in Fall 2020) is limited. The small number of beyond-first-year students in our data is also a consequence of beyond-first-year lab courses being offered less frequently than first-year lab courses, perhaps not even every year and, therefore, failing our selection criterion of having a matching course in 2019. A qualitative analysis of the impact of emergency remote teaching on beyond-first-year courses will be the subject of future work.

As we are interested in the general trends in student views in 2020, we amalgamate student responses from all the identified courses. To answer RQ1, we consider the total E-CLASS score as a measure of student views. For the analysis, we compare the distributions of pretest scores across years to establish the degree of similarity in student responses at the beginning of their course. Then, we compare the distributions of post-test scores across years to identify any differences. In addition, we consider the distributions of the shift in scores from pretest to post-test (specifically, for each student subtracting their pretest score from their post-test score). The shift is a useful measure because (i) it accounts for a students' pretest score, and, hence, (ii) it tells us directly the effect of the period of time the student is in the course.

To perform these comparisons, we first look at summary statistics (means, medians, etc.), then use the Mann-Whitney U test [25] to establish the statistical significance of any differences in the distributions of the E-CLASS scores from two groups of students. We use this test because the distributions we are considering are categorical and non-Gaussian. Additionally, we choose to report effect sizes (through the rank-biserial r [26]) and 95% confidence intervals as our measure of practical significance to provide easy to interpret measures independent of sample size as recommended by Cohen [27] and more recently Rodriguez [28]. We report the magnitude of r and, as such, it is constrained to have a value in the range [0, 1]. If all student scores from the two groups are ranked in order, then $r = 0$ corresponds to there being an equal likelihood for the next score to be from either of the two groups, while $r = 1$ corresponds to all the scores in one group being higher (or lower) than all scores in the second group [29]. These effect size estimates are generated using the *wilcoxonR* function of the statistical programming language *R* [30,31]. In interpreting the values of the effect size, we follow the guidance of McGrath and Meyer [32], where values of $r = 0.10$, $r = 0.24$, and $r = 0.37$ correspond to small (merely statistical), medium (subtle), and large (obvious) effect sizes respectively (for the relation to Cohen's d [33] see Table 7 in Ref. [34]).

Previous work has shown that there is a strong dependence of post-test E-CLASS score on the pretest E-CLASS score [35]. Furthermore, other variables that can be predictors of post-test score include the choice of major (physics or nonphysics) and gender. Therefore, in addition to the above analysis, we model the post-test score with a linear model (ANCOVA) including a categorical variable

for the year to identify and isolate the effect of students taking a lab class in 2020.

To answer RQ2, we consider each of the 30 questions from the E-CLASS separately. We calculate the mean score (which exists in the range [-1,1]) and compare the shift in mean score between the two years. A student's score for each question is in the discrete set $\{-1, 0, 1\}$ and, thus, the shift in score (post-test score minus the pretest score) is in the discrete set $\{-2, -1, 0, 1, 2\}$. Hence, we again make use of the Mann-Whitney U test, as described above, to determine the significance of the differences between distributions of the shift in score for different groups of students (e.g., from one year to the next). Additionally, we report the p values, p, associated with the Mann-Whitney U tests for each item to demonstrate to what extent we are able to draw statistically relevant conclusions based on an analysis of changes in individual item scores. That is, we will report on individual items that do not satisfy the null hypothesis that the distributions of the shifts in score are the same between two years at the critical p level of $\alpha = 0.01$. We use the Holm-Bonferroni correction to avoid type I errors (false positives) [36–38], meaning that the effective p level to consider the first (lowest p value) of 30 items as not a false positive is $p < 3 \times 10^{-4}$.

C. Course contexts and demographics

In our data, there are 14 courses in which E-CLASS was used in both Spring 2020 and Spring 2019. As many instructors teach, or administer, multiple courses, these 14 courses come from a sample of eight institutions. Of these, one is a master's awarding institution, and seven are Ph.D. granting institutions [39]. In the fall data, there are 20 courses in which E-CLASS was used in both 2020 and 2019, which come from a sample of 14 institutions. Of these, four are 4-year colleges, two are master's awarding institutions, and eight are Ph.D. granting institutions.

In the Spring 2020 data, the minimum number of students in a course that we have complete data for is 9 and the maximum is 454 with a median number of students of 95. In the Fall 2020 data, the minimum number of students in a course that we have complete data for is 4 and the maximum is 358, with a median of 55. We emphasize that these numbers are only a conservative estimate for the numbers of students in a course, as not all students would have participated in the E-CLASS (see Sec. II D for more details). The decrease in the number of students per course between Spring and Fall 2020 may be a result of lower enrollment, but our data are not suitable to draw any conclusion on that, as only 9 of the 20 courses in Fall are

TABLE I. Student demographic information from spring and fall for introductory courses in the years 2018, 2019, and 2020. Courses in 2018 are included only if the same course occurred in 2020 and 2019. Race and ethnicity labels correspond to those used by the U.S. Census Bureau [40]. The category "Physics" major includes physics and engineering physics majors only. "Engineering" includes all engineering other than engineering physics. "Other Science" majors include all other science majors (biology, chemistry, math, computer science, etc.) not included in the physics category, "Nonscience" majors include all other majors as well as undeclared majors.

	Spring			Fall		
	2018	2019	2020	2018	2019	2020
Number of courses	7	14	14	16	20	20
Number of students	1082	1632	1700	1934	1983	1918
Female	40.7	48.0	55.6	52.3	50.0	53.6
Male	59.3	52.0	44.4	47.7	50.0	46.4
American Indian or Alaska Native	0.9	0.7	1.1	0.7	1.1	1.2
Asian	10.2	12.6	12.9	15.9	18.5	17.8
Black or African American	11.6	10.5	10.2	10.0	9.4	10.6
Hispanic or Latino	9.1	6.2	6.5	7.2	8.5	13.5
Native Hawaiian or other Pacific Islander	0.6	0.3	0.4	0.4	0.5	0.4
White	67.7	70.5	70.4	66.9	63.9	60.8
Other ethnicity	2.5	1.6	2.6	2.4	2.5	2.1
Unknown ethnicity	3.8	3.9	2.8	3.7	4.7	3.8
Physics	6.5	6.1	5.2	3.9	5.5	3.8
Engineering	23.6	26.2	24.9	19.5	21.6	24.1
Other Science	48.2	40.7	42.7	45.5	41.2	42.2
Nonscience	21.8	27.0	27.2	31.0	31.7	29.9
1st year	28.7	35.5	32.9	16.3	19.2	20.8
2nd year	32.7	27.1	27.3	30.2	31.7	32.0
3rd year	26.6	24.5	27.9	35.0	33.1	31.1
4th year	9.4	10.4	10.6	15.2	12.9	14.0
5th year or higher	2.6	2.5	1.3	3.3	3.1	2.1

the same as those in Spring. The large range of number of students in each course poses the risk of the courses with larger populations dominating the statistics of our results. However, we note that the within course variation of student scores is larger than the variation of scores between courses. Specifically, the standard deviation of the mean scores for each course (2.8 points in Spring 2020 and 2.6 points in Fall 2020) is less than half of the mean of the within course standard deviation of scores (7.4 points in Spring 2020 and 6.9 points in Fall 2020), suggesting that amalgamating all student scores is reasonable. Nevertheless, we remain cognizant of this fact and will address it again when presenting the results.

A full breakdown of student demographics (gender, race or ethnicity, major, and student year) is provided in Table I. We see that the demographic characteristics of the students in the samples from 2020 and 2019 are very similar in both spring and fall semesters. We have also included demographic information for the 2018 semesters, to help justify our later comparisons. In Spring 2018, there are half the number of courses than in the 2019–2020 data, and the resulting demographics are slightly different. As such, comparing 2018 with 2019 it would not be unexpected to find differences in E-CLASS scores (with all other things remaining equal) than when comparing 2019 with 2020 just because of the underlying differences between the distributions of student backgrounds. Nevertheless, this can be used as a conservative upper estimate of the size of difference we might expect to see, and, therefore remains a useful comparison to make. In Fall 2018, the dataset contains two-thirds of the courses that are in the 2019–2020 sample, while actually having more student responses than in 2020. Consequently, the demographic data is very similar in 2018 to 2019–2020, and we can have greater confidence in performing a comparison between 2018 and 2019 E-CLASS scores to establish the expected size of changes between two pre-pandemic years.

D. Limitations of the study

A major limitation of this work is the scope in which the results can be generalized. This is most noticeable when considering the demographics of the student population we have sampled. We draw a comparison between the distribution of the race or ethnicity of students presented in Table I and the recent work of Kanim and Cid [41] indicating that the student population in our sample overrepresents White students compared to the population of students who took the SAT in 2015. Furthermore, the majority of higher-education institutions in the study are Ph.D. granting institutions, which is not representative of the general educational landscape in the U.S.

Another source of bias in our data is that all the instructors have demonstrated a sustained interest in understanding how well their students are learning by choosing to use the E-CLASS. The sustained nature of this interest is specific to this study, as we have deliberately selected courses in which the E-CLASS has been used for at least two years, and its use has been maintained during the pandemic. Through this statement, we are not implying that instructors who did not continue to use the E-CLASS were not interested in learning about their students' views of experimental physics, and we recognize that the extra workload for both instructors to administer and students to participate in the E-CLASS is a legitimate reason to end participation. A corollary of this point is that courses where the pandemic had a larger impact on either the instructors and/or students may not be present in our data, which echoes the first limitation discussed on how these conclusions should not be generalized. This bias is one area we will return to in Sec. IV.

While we have explicitly stated that we are not investigating specific courses and how they were adapted in 2020, the shear variety of approaches taken by instructors when faced with the challenge of adapting their courses (see, for example, Ref. [8]), will mean that our analysis will be sensitive to only "global" effects that would be correlated across the sampled population. We mention this here as it may be considered a limitation, but is also a deliberate feature of our study's design and relates directly to our research questions.

A more specific limitation of this work is that, by definition, we do not sample students who dropped out of the course or who chose not to complete the post-test. This is due to the requirement of having matched pretest and post-test responses. Inspection of the number of pretest and post-test responses before performing matching shows a decrease in responses from pretest to post-test of 20% in Spring 2020 and a decrease of 24% in Fall 2020, which are comparable to the decreases in participation seen in 2019 (for spring there was a 28% decrease and for fall there was a 22% decrease). Interestingly, the magnitude of the decrease in responses was lowest in Spring 2020, perhaps due to the fact that almost all teaching was being conducted online at the time and, thus, students may have found it easier to complete an online survey.

The fact that the E-CLASS surveys are administered online provides significant advantages for instructors, researchers, and many students, but we also note that this may disadvantage some groups of students. This has been particularly highlighted during 2020, with access to a reliable internet connection to conduct remote learning, as well as access to computers as and when they are needed, proving a problem for a large minority of students in the U.S [8,42]. It is, therefore, important to remain aware of these issues when considering the conclusions of this work.

E. Ethical considerations

Whether or not to conduct research during a pandemic is an ethical question that we considered before proceeding

with collecting data for this research. This includes, but goes beyond, the normal institutional review board (IRB) approval. The concerns raised included whether it would be ethical to study a population who may be suffering from the health implications of COVID-19, have to care for relatives who contracted the virus, have increased childcare responsibilities due to school closures, or have to deal with the economic and psychological consequences of the pandemic. We decided on the following guiding principles for the study, some of which are standard in all of our studies and in our IRB protocol. Participation in this research has been optional. Furthermore, we provided the opportunity for instructors to opt out in Spring 2020 (as the pretest surveys had already been administered) and to opt-in during the Fall 2020 data collection to receive the remote lab survey questions. For students, who would normally have completed the E-CLASS as an integrated part of their course, these remote lab survey questions were not mandatory for gaining the course credit for completion, if offered by the instructor. Furthermore, when designing these additional questions we remained conscious of the added time they would take for the students to complete and made decisions on which questions to include to minimize this time.

We also considered the potential benefits of continuing data collection using the E-CLASS during the pandemic. Specifically, by documenting student and instructor experiences, we can recognize the resilience and creativity of those participating in the research, thereby showing that the participants' views and experiences are valued by the community. Through understanding the impacts of remote lab instruction during the pandemic, instructors and researchers can learn from both the successes and the challenges faced to inform future instructional choices within and beyond the confines of this pandemic, thus benefiting all students in physics lab courses. The longer term benefits of effective remote instruction arise from the potential to identify techniques that facilitate the participation of students with diverse accessibility needs as well as nontraditional and rural students in physics courses. Therefore, in balance, we believe that these benefits far outweigh the risks, especially when considering that the burden of participation in the research is minimal.

III. RESULTS

A. Spring 2020

Because of the rapid spread of COVID-19 during Spring 2020, many higher-education institutions in the U.S. switched to remote learning. This happened on, or around, March 16th, approximately halfway through most spring semesters (of the 14 courses we consider, the mean number of days remaining of the course corresponded to 46% of the total course duration). Instructors had to swiftly transition to teaching online; in some cases only having a few days to do so.

We describe the types of activities that instructors resorted to during this period, in order to illustrate to the reader the instructional practices that correspond to the following results that we present. From the sample of courses from Spring 2020 featured in this research, in 79% of the courses, the instructor started to provide videos of themselves performing an experiment; in 64% of courses, the instructor provided students with data for them to analyze; in 43% of courses, instructors started to use online simulation tools; in 14% of courses, students were tasked with collecting data at home; and in 7% of courses, students reported completing less written work, such as lab reports (see Appendix A in the Supplemental Material [20] for the analysis used to generate these numbers). Students from all courses reported a shift from mostly group work during in-person instruction to individual work while working remotely.

1. Comparison of E-CLASS score distributions (RQ1)

Having established in Sec. II C that the student demographic details in Spring 2020 and Spring 2019 are equivalent, we look to compare E-CLASS scores between these two groups. First, the pretest distributions [left-hand plot in Fig. 1(a)] are remarkably similar between the two

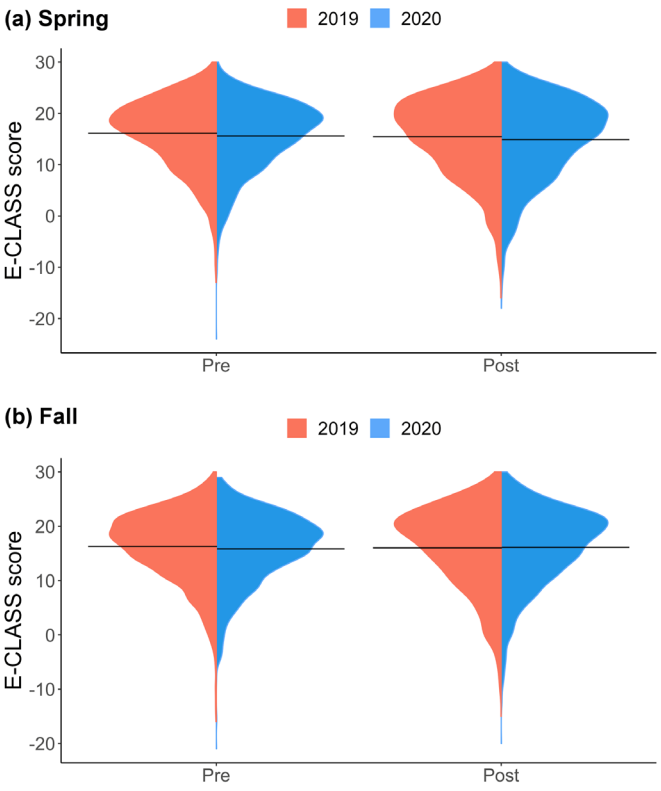

FIG. 1. Violin plots showing the distribution of pretest (left plot) and post-test (right plot) E-CLASS scores for all courses that occurred in (a) Spring 2019 and Spring 2020, and (b) Fall 2019 and Fall 2020. Horizontal black lines indicate the mean score for each distribution.

TABLE II. Summary statistics for the distributions shown in Fig. 1 and for the equivalent distributions in 2018. Δ corresponds to the post-test score (Post) minus the pretest score (Pre). Uncertainties on the mean scores are the standard error. The effect size (r) and the 95% confidence intervals are calculated using the result of the Wilcoxon signed-rank test [43] using the *wilcoxonPairedR* function [31] and is used here because we are comparing matched data from the same student from pretest to post-test.

	Mean				Q1			Median			Q3		
Spring	Pre	Post	Δ	Effect size (r)	Pre	Post	Δ	Pre	Post	Δ	Pre	Post	Δ
2018	17.2 ± 0.2	16.3 ± 0.2	−0.9	0.14 ± 0.06	14	12	−2	18	18	0	22	22	0
2019	16.1 ± 0.2	15.5 ± 0.2	−0.7	0.08 ± 0.05	12	11	−1	17	17	0	21	21	0
2020	15.6 ± 0.2	14.9 ± 0.2	−0.7	0.09 ± 0.05	12	10	−2	17	16	−1	21	21	0
Fall													
2018	16.1 ± 0.2	15.3 ± 0.2	−0.8	0.11 ± 0.04	12	11	−1	17	17	0	21	21	0
2019	16.3 ± 0.1	16.0 ± 0.2	−0.3	0.02 ± 0.04	13	11	−2	17	18	+1	21	22	+1
2020	15.8 ± 0.1	16.1 ± 0.2	+0.3	0.06 ± 0.05	12	12	0	17	17	0	20	21	+1

years, with the same first, second, and third quartile scores (Table II). The mean pretest score in 2020 is 0.5 points lower than in 2019, but care must be used when comparing means, as the distributions are clearly not Gaussian. To evaluate the null hypothesis that the two pretest distributions (Spring 2020 and Spring 2019) come from the same underlying distribution, we use the Mann-Whitney U test [25] and find the effect size to be not significant (negligible) as shown in Table III.

The fact that the two pretest distributions are so similar is useful, as then we can have more confidence that any difference in the post-test scores in Spring 2020 compared with Spring 2019 may be attributed to the impact of the transition to emergency remote instruction (especially given that in Spring 2020 the pretest was taken under pre-pandemic circumstances). Nevertheless, we find that the post-test distributions [right-hand plot in Fig. 1(a)] are also similar between Spring 2019 and 2020, though in 2020 the first and second quartile post-test scores are slightly lower than the Spring 2019 post-test scores (Table II). The mean post-test score in Spring 2020 is 0.6 points lower than in 2019, while the Mann-Whitney U tests shows a negligible effect size (Table III).

TABLE III. Effect sizes calculated from the results of Mann-Whitney U tests comparing the distributions of student scores (cf. Fig. 1) between pretests and post-tests for different years. Uncertainties correspond to an estimate of the 95% confidence interval.

		Year comparisons		
Spring		r (Pre or Pre)	r (Post or Post)	r Post-Pre
2018	2019	0.07 ± 0.04	0.05 ± 0.04	0.03 ± 0.04
2019	2020	0.04 ± 0.03	0.03 ± 0.03	0.00 ± 0.04
Fall				
2018	2019	0.01 ± 0.03	0.05 ± 0.03	0.05 ± 0.03
2019	2020	0.04 ± 0.03	0.00 ± 0.03	0.04 ± 0.03

Comparing the distributions of the shifts in E-CLASS scores between Spring 2019 and Spring 2020 with the Mann-Whitney U test also gives a negligible effect size. This further supports the result that there is no statistical difference between student total E-CLASS scores in Spring 2020 compared to Spring 2019. We can establish the validity of our interpretation of these effect size estimates by comparing them with the effect sizes when comparing Spring 2018 with Spring 2019. We see in Table III that the effect sizes for the Spring 2019–2020 comparisons are all smaller than those for the Spring 2018–2019 comparisons.

In order to isolate whether changes between 2019 and 2020 exist, independent of the known effects of pretest score, gender, and major, in Appendix B [20], we have constructed linear models to describe the post-test E-CLASS score as a function of the continuous variable: the pretest E-CLASS score; and the categorical variables: year in which the E-CLASS was taken, major, and gender. Even when accounting for these variables, the results indicate that there is no significant difference between post-test E-CLASS scores in Spring 2020 compared with Spring 2019 and, hence, supporting the results reported above. Furthermore, the only variable with a non-negligible effect size was the pre-test score, suggesting that in addition to the calendar year, both major and gender are not predictors within this sample of students.

2. Comparison of individual E-CLASS question scores (RQ2)

While we observe no differences in overall E-CLASS scores between Spring 2019 and Spring 2020, it is also possible that student responses to individual questions in E-CLASS changed between the two years in such a way that the overall score did not change. We consider the difference in score between the post-test and the pretest here to account for the fact that, for some questions, the pretest scores in Spring 2020 were different from Spring 2019 (maximum effect size from the Mann-Whitney U test comparing the distributions of scores in the pretest

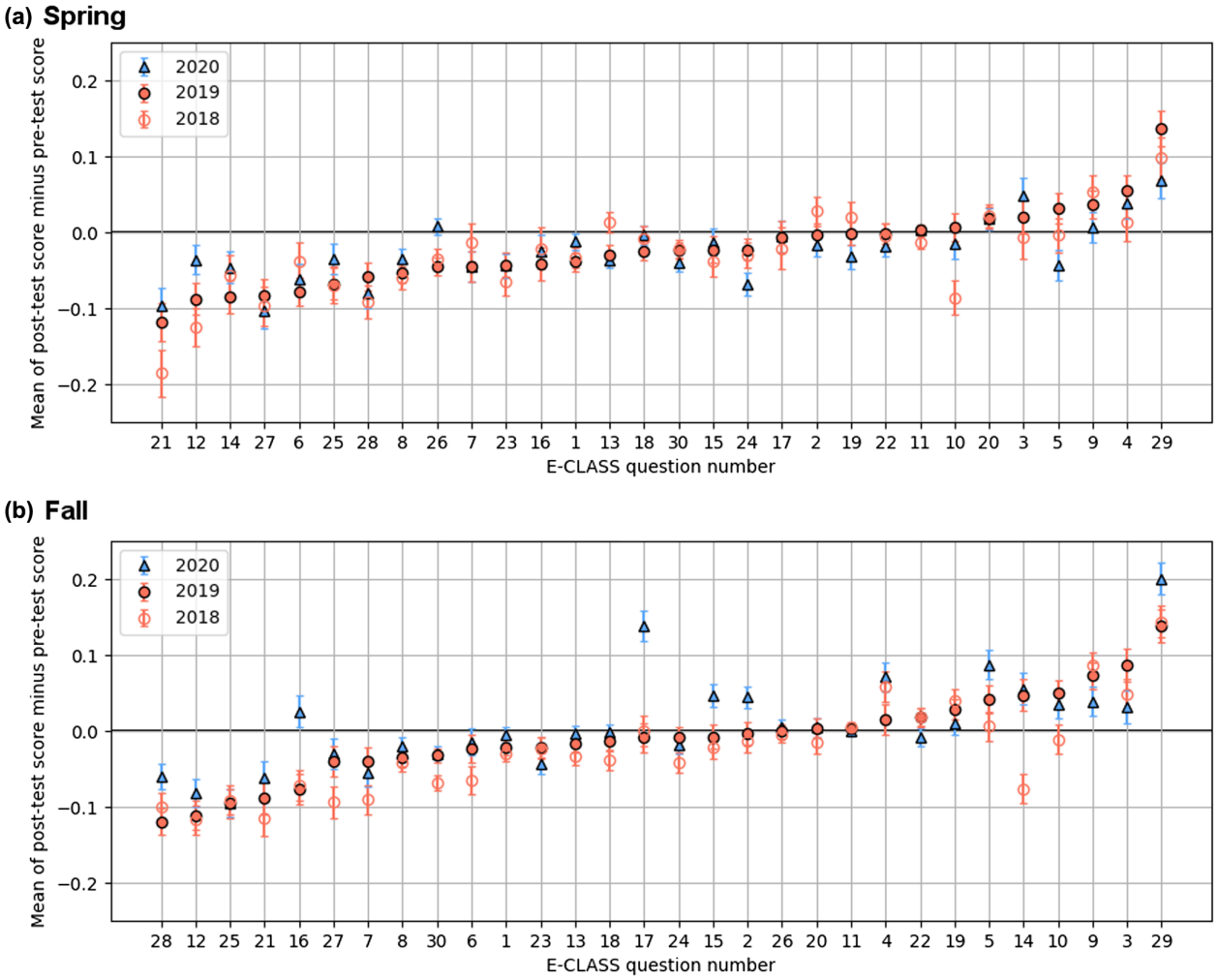

FIG. 2. Mean of the differences in student E-CLASS scores (post-test score minus pretest score) for individual questions for (a) spring and (b) fall semesters. In each panel blue triangles represent the 2020 data, solid red circles the 2019 data, and open red circles the 2018 data. Error bars indicate the standard error on the mean. A list of the E-CLASS questions can be found in Ref. [44]. Items are ordered based on increasing 2019 values. Apparent missing symbols for some questions indicate that the scores are the same.

questions between the two years is $r = 0.06 \pm 0.03$). For reference, we provide tables of mean pretest and post-test scores for both Spring 2020 and Spring 2019, along with effect size values in Table VII in Appendix C in the Supplemental Material [20].

In Fig. 2(a), we see that there is some variation in the shift between pretest and post-test item scores between 2020 and 2019, however, the effect size calculated from the Mann-Whitney U test comparing the distributions of the shifts (post-test score minus pretest score) for each of the differences between years for each item is negligible ($r \leq 0.05$ and the most significant p value, $p = 3 \times 10^{-3}$, is not small enough to reject the null hypothesis; see Table VII). We have included the shift from pretest to post-test question scores in Fig. 2(a) for Spring 2018 data to help illustrate what pre-pandemic year-to-year variation may lead to in E-CLASS scores. It then becomes clear that changes in scores between 2019 and 2020 are no greater than the variation expected between years before the pandemic and hence validating the interpretation of the effect size measures being negligible.

B. Fall 2020

In Fall 2020, most institutions had decided on their modality for instruction based on local circumstances. The possible different modalities, which we surveyed for, were entirely in person, entirely remote, hybrid of in-person and remote, and a mixed modality that changed during the semester. Of the 20 courses in our sample for Fall 2020, we found, based on student reporting, that 90% of these courses were entirely remote (see Appendix A [20] for details of the calculation).

It is clear that many instructors deliberately tried to address some of the challenges of teaching a remote lab while maintaining aspects of the in-person experience they valued, especially those relating to group work and

providing students with hands-on activities [8]. For instance, in 65% of the courses in Fall 2020, students reported that they often or always worked in groups, and in 60% of the courses, students reported they often or always completed hands-on activities. For all students reporting completing hands-on activities (including those who only rarely, or sometimes did hands-on activities), we asked follow-up questions to ascertain the nature of these activities (see Appendix A [20] for these questions): in 50% of courses, students reported using household materials (such as cardboard and tape); in 35%, students used smartphones to record data; in 25%, students reported having to purchase equipment; and in 20%, students reported equipment was given to them by their institution.

Other activities students reported participating in during Fall 2020 were written activities, such as lab reports or proposals (students in 80% of courses reported often or always doing this activity), using lab notebooks (40% of courses), using simulations (35% of courses), and watching videos (25% of courses). In 85% of courses, students reported using simulations to perform measurements (even if only rarely or sometimes). In 65% of courses, students reported collecting data from videos (either from instructors performing the experiment live or using video analysis techniques).

1. Comparison of E-CLASS score distributions (RQ1)

Following a similar line of analysis as in Sec. III A 1 we first identify that the shapes of the distributions of E-CLASS scores in Fall 2020 are similar to Fall 2019 [Fig. 1(b)]. While the first, second, and third quartiles of the pretest E-CLASS score distributions are not exactly the same when comparing Fall 2020 with Fall 2019, they differ by at most 1 point between the years (Table II). The mean pretest E-CLASS score in Fall 2020 is 0.5 points below the mean pretest E-CLASS score in Fall 2019. To evaluate the null hypothesis that the two pre-test distributions (Fall 2020 and Fall 2019) come from the same underlying distribution, we use the Mann-Whitney U test and find a negligible effect size (Table III).

The mean post-test score in Fall 2020 is 0.1 points higher than in Fall 2019, corresponding to a negligible effect size when comparing the distributions of post-test scores using the Mann-Whitney U test (Table III). This mean post-test score is also the only mean score of any semester that is higher than the mean pretest score, though this positive shift also has a negligible effect size when comparing the distributions of shifts using the Mann-Whitney U test. Inspection of the quartile values in Table II suggests that this positive shift in the mean, rather than being due to an increase in all students' scores, is due to the absence of a drop (from pretest to post-test) in the first quartile score which is otherwise seen in all other semesters.

Comparing the effect sizes of the post-test minus pretest shift in student score between Fall 2020 and Fall 2019, with those comparing Fall 2019 with Fall 2018, we see that the sizes of the changes in Fall 2020 are similar in size to when comparing year-on-year changes pre-pandemic. This again validates our interpretation of these effect sizes as being negligible.

In Appendix B [20], we construct a linear model of the post-test score using the pretest score, the choice of major (physics or nonphysics), gender, and year (2020 or 2019) as predictor variables. When accounting for all these factors, we find that the effect size of the variable representing the year is negligible. Indeed, and similarly to the spring results, the only variable with any non-negligible effect size is the pretest score.

2. Comparison of individual E-CLASS question scores (RQ2)

Similar to Spring 2020, in Fall 2020 the majority of individual questions show little difference in mean scores compared to previous years [Fig. 2(b)]. However, some questions show a marked increase in the size of the shift between pre-test and post-test compared to both Fall 2019 and Fall 2018. We discuss only the two largest changes, as these two questions show statistically significant differences (at the 1% Holm-Bonferroni corrected level) and have the largest effect sizes of all items. For the interested reader, we provide extended data in Appendix C [20].

The largest difference between the years is for the statement associated with question 17:

> *(Q17) When I encounter difficulties in the lab, my first step is to ask an expert, like the instructor.*

At the end of their course in Fall 2020, more students changed their view to disagreeing with the statement (the more expertlike response) compared to the start of their course. The shift shown in Fig. 2(b) for Q17 has the largest effect size of any comparison between 2020 and 2019 ($r = 0.08 \pm 0.03$, $p = 2 \times 10^{-7}$). This result seems perfectly reasonable, given that 90% of the courses were entirely remote, and, therefore, students would not have been in the same room at the same time as the instructor and could not necessarily simply talk to the instructor when they needed help. However, whether it truly reflects a shift toward an expertlike attitude is not clear, as students in a remote lab may have interpreted the statement prompt differently compared to students in in-person labs for which the statement was designed. Specifically, the expertlike response is to disagree with the statement of Q17, with the implicit meaning that an expert's first step would be to try to troubleshoot or solve the problem themself, before looking for outside help. For some students, this indeed might have been the case in the remote environment, not out of choice but necessity. However, interpreting this result as such is tenuous, as (i) including "instructor" as an

example of an expert could prompt students to think more about how easy it was to communicate with their instructor, rather than what they did when they encountered difficulties; and, (ii) because of the different context from that which the survey question was created, we cannot be certain as to what, if any, other sources or actions the students took for their first step.

The second largest change in response, in terms of both the difference in size of the pretest to post-test shift and effect size ($r = 0.06 \pm 0.03$, $p = 7 \times 10^{-5}$), is in the statement associated with question 16:

(Q16) The primary purpose of doing a physics experiment is to confirm previously known results.

Again, the expertlike view is to disagree with this statement. In previous years, responses to this question showed a mean negative shift from pretest to post-test, while in Fall 2020 there is a small positive shift. Negative shifts in E-CLASS scores have previously been associated with "guided" labs [21], which often use confirmatory experiments. The change in student responses (or rather lack of a drop in score) seen for Q16 could suggest that the lab activities during Fall 2020 were more open ended and less confirmatory than in previous semesters. This is supported by some anecdotal evidence from Spring 2020 on the type of activities instructors switched to when transitioning to emergency remote instruction due to limitations in the equipment available to students in their homes [8].

As mentioned in Sec. II D, our analysis deliberately does not distinguish between different courses. However, as some courses have more students than others, this may bias the previously discussed results such that it is not possible to claim that this effect was seen generally across the courses in our sample. By considering the distribution of the mean scores for each course on Q16 and Q17, we see that it is indeed the case that the positive shift from pretest to post-test occurred in the majority of courses (Fig. 3). For Q16, this was the case in over 50% of the courses, while in Q17 over 75% of the courses saw a positive shift.

To close the presentation of results, we make a brief comment on how we can see a change in individual question scores and not in the total E-CLASS score. Inspection of Fig. 2(b) shows that while some questions, as we have discussed, show positive shifts, a number of others show small (i.e., not significant $r \leq 0.05$, $p > 0.01$) negative shifts (in comparison to previous years), which account for the overall balance in the total score. It is interesting to note that those questions with small negative shifts are related to actually performing experiments, for example, the statement of question 9 is

(Q9) When I approach a new piece of lab equipment, I feel confident I can learn how to use it well enough for my purposes.

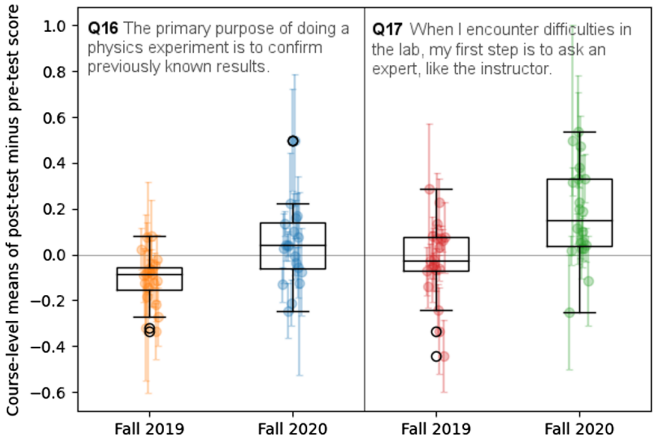

FIG. 3. Box and whisker plots illustrating the distribution of student post-test minus pretest scores on questions 16 and 17 (left and right boxes respectively) in Fall 2019 (left) and Fall 2020 (right). These items are those from Fig. 2(b) that show the largest effect size in 2020 compared to 2019. The center, horizontal line of each box corresponds to the median student difference in scores, while the lower and upper edges of the box correspond to the first and third quartiles, respectively. The whisker length is the furthest data point within 1.5 times the interquartile range from the edge of the box. Data points outside this range are indicated with open black circles. Overlaid on each box and whisker plot is plotted, as colored circles, the mean score for each course, with error bars indicating the standard error of the mean.

A negative shift here seems plausible if students do not have access to what they think counts as "lab equipment." Conversely, some questions with small (not significant) positive shifts seem more related to student affect, for example, the statement from question 2:

(Q2) If I wanted to, I think I could be good at doing research.

This could be a consequence of students having to perform experiments in a more independent environment than in previous years. As the effect size of these shifts compared to previous years is negligible, and may be course dependent, investigation of them falls outside of the scope of the present work. We will use our results to answer the research questions, discuss the implications of our results, and compile the outstanding questions raised by them in the following section.

IV. DISCUSSION AND CONCLUSIONS

The results of Sec. III A 1 and Sec. III B 1 allow us to answer the first research question. For the sample of courses that have been included in our study, there was no significant difference between the distributions of total E-CLASS scores in both Spring and Fall semesters in 2020 compared to the same semesters in 2019, indicating that, in general, students neither became more nor less expertlike in their views about experimental physics in their lab classes

in 2020 compared to previous years. Hence, we can provide no evidence, at the level of the total E-CLASS score, to support our hypothesis that student views about experimental physics would become less expertlike during emergency remote instruction. While the nature of our analysis prevents us from making causal inferences, in the first part of this section we enumerate possible explanations for the observed, and what may be considered surprising, consistency of total E-CLASS scores in 2020.

First, while the answer to RQ1 is the same for both spring and fall semesters, we emphasize that the context of these two semesters was very different. In Spring 2020, courses suddenly transitioned to remote instruction, with over 50% of the course having already been completed in person, while in Fall 2020, not only were most courses in our sample entirely remote, instructors had more time to prepare their courses (compared to a few days or, at most, weeks in spring). Therefore, it may be argued that the impact of the curtailment of in-person, hands-on lab and group work on student attitudinal development in spring was limited by the fact that students still experienced the majority of their course as planned. In contrast, in fall, the necessary imposition of remote instruction led to instructors having to think creatively about how they may achieve the learning goals of their courses in this new environment. Therefore, students would not have experienced a "normal" physics lab and it would be reasonable to assume that this would have affected the development of their views toward experimental physics. The fact that this has not been the case simply highlights the extraordinary efforts of instructors in our sample in adapting old and preparing new course material for the fall semester. Though, we note this explanation does relate to one of the aforementioned limitations of our study (Sec. II D) in that the instructors participating in our research have demonstrated a sustained interest and engagement with the physics education community, meaning it may not be possible to generalize these results.

Another possible explanation is that students hold robust views that are developed over a lifetime of experiences, and, therefore, the impact of one course in one semester on those views is likely to be small (see, for example, the effect size when comparing pretest and post-test distributions in Table III). However, by design, the E-CLASS asks students: "What do YOU think when doing experiments for class?" This deliberately asks students to think about their present class and not experimental physics in general, therefore favoring our primary explanation of these results described in the preceding paragraph.

A third explanation involves the possibility that what we have observed is a selection effect. In both semesters in 2020, there were numerous pandemic related pressures that might cause students to drop out of their course (or not enroll in the first place for the fall semester). These students are excluded from the analysis of the E-CLASS and so we are unable to measure the development of their views. However, there are a number of reasons that this explanation may be discounted. First, lack of participation is not just an issue for the data from students in 2020, but also in previous years. Second, student numbers responding to the survey (before matching of pretest and post-test responses) in both Spring and Fall 2020 are similar to 2019, with similar drop-out rates (see Sec. II D). Third, the composition of the cohorts in terms of majors and degree year in each semester are comparable (see Table I), suggesting that there were no large shifts in the populations of students taking the courses in our sample in 2020 compared to 2019. Nevertheless, we also believe it to be important that we remind the reader of the underrepresentation of minority groups within our sample. In the balance of evidence, it seems the most likely reasons for the sustained development of student views toward experimental physics labs during 2020 is a result of the ingenuity of the lab instructors who successfully adapted their courses as well as the facility of the students who responded to the E-CLASS to continue their studies despite the circumstances (recognizing that this was not possible for all students).

The results in Sec. III A 2 and Sec. III B 2 allow us to answer the second research question. In spring, we found that no question on the E-CLASS survey showed a significant difference from the previous two years. Thus, we can say that no specific views about experimental physics changed in the spring semester of 2020 more or less than in Spring 2019, again providing no evidence to support our hypothesis. This reinforces our explanation of the spring results for the first research question, as this explanation was founded on the idea that students gained enough experience in the first half of the semester that meant their views about experimental physics related to the course had already been established.

In fall, we found that expectations around where to find help and support during a lab course changed, with students relying less on their instructor; a possibly unsurprising consequence of remote instruction and, therefore, to some extent, outside of the instructors control. Whether students actually held a more expertlike view, such as attempting to solve a problem themselves first before turning to an expert, or were simply responding to the prompt with the view that they could not contact their instructor remains an open question.

We also found some slight evidence that students' views on the purpose of experimental physics became more expertlike, contradicting our hypothesis in this one respect. This attitudinal change further supports our explanation for the answer to the first research question, as this suggests that instructors managed to design effective activities that provided students opportunities to understand that the role of experimental physics is not just to confirm previously known results. This issue has been highlighted in the research literature for a number of years [45] and as higher-education institutions were forced in 2020 to allow changes to course content and delivery, perhaps instructors

took the opportunity to enact such changes. Whether these changes will remain in place once in-person labs resume remains to be seen.

Returning to the situated theory on which our hypothesis was based, we cannot argue that simply being in a teaching lab is necessary for the development of students' views about experimental physics [46]. Rather, our results suggest that the type of activities students undertake within a "lab class," whatever form that may be, are what provide the context for students' own reflections to take place. Specifically, and importantly, these results cannot be used to imply that activities that include group work and hands-on data collection are not required for the development of student views about experimental physics, as, in a majority of the courses sampled from Fall 2020, students reported often or always having activities with either of these two aspects.

We finish by outlining the questions that have arisen during this study. One important question is what instructional changes were most successful in terms of impact on E-CLASS score? Or, put another way: why the remote labs did not alter student views of experimental physics compared to in-person classes? This is of particular interest, as when labs return to in-person instruction there might be aspects of teaching that would be useful to retain. For example, from our results as discussed above, we may ask what specific activities did instructors employ that led to students' views on the purpose of experimental physics to change?

Another question is how did student experiences of emergency remote instruction influence their views as seen in their E-CLASS score? For example, did students who faced challenges with group work show lower scores in the E-CLASS, and in which specific items? To answer these questions we plan to conduct further quantitative and qualitative analysis on the student experience of physics labs during the pandemic, to explore these topics in more depth.

The final question that we wish to raise is what was the impact on student views in beyond-first-year courses? This is interesting because students in these courses are generally physics majors and have higher pretest scores on the E-CLASS than the first-year courses presented in this work. Furthermore, these courses often rely on more specialized experimental apparatus that may not as easily be provided to students remotely. Therefore, the experiences of these students during 2020 in terms of their lab classes might be notably different from the courses we have discussed. This is an important limitation of remote instruction to recognize, as labs provide an opportunity for students (in both first-year and beyond-first-year courses) to learn how to use scientific equipment they may have never seen before and to do so in a safe environment.

In conclusion, we have seen that, in our sample, student expectations and epistemologies about experimental physics (their *views*) were not significantly impacted by the transition to emergency remote instruction in Spring 2020 nor by the continuation of remote instruction into the Fall of 2020. Therefore, we did not find evidence to support our hypothesis that student views toward experimental physics would deteriorate, over the period of one semester, as a result of emergency remote instruction. This conclusion has major caveats, which have been detailed in Sec. II D, including the issue of representation, both with respect to marginalized groups within physics, and those to whom the COVID-19 pandemic had the greatest impact. Therefore, we do not attempt to claim that the results can be generalized. Given the upheavals to education that took place in 2020, and with no intention to minimize the potential suffering of students and instructors, we believe that maintaining the same outcomes as in a year without a pandemic (2019), along this one dimension of student views, can be considered a relative success for both instructors and students in our dataset. Furthermore, small improvements seen in Fall 2020 in the score for individual E-CLASS questions regarding access to experts and the purpose of experimental physics suggest that further qualitative research is needed to understand these changes. Especially, further investigation is required into the necessary negative shifts in other E-CLASS questions (being individually not significant) that compensate for these observed positive shifts to produce no overall change in the total E-CLASS score.

ACKNOWLEDGMENTS

We would like to thank all the instructors and students who participated in this research study, especially during this turbulent time. We want to thank Mary-Ellen Phillips for some early analysis, Nidhal Sulaiman for discussions on analysis methods, and Bethany Wilcox for comments on the manuscript. This work is supported by NSF Grants No. DUE-2027582 and No. PHY-1734006.

[1] E. Dorn, B. Hancock, J. Sarakatsannis, and E. Viruleg, *COVID-19 and Student Learning in the United States: The Hurt Could Last a Lifetime* (McKinsey & Company, 2020), https://www.mckinsey.com/industries/public-and-social-sector/our-insights/covid-19-and-student-learning-in-the-united-states-the-hurt-could-last-a-lifetime.

[2] M. Kuhfeld, J. Soland, B. Tarasawa, A. Johnson, E. Ruzek, and J. Liu, Projecting the potential impact of COVID-19

school closures on academic achievement, Educ. Res. **49**, 549 (2020).
[3] P. Engzell, A. Frey, and M. D. Verhagen, Learning inequality during the COVID-19 pandemic, SocArXiv ve4z7, Center for Open Science, https://ideas.repec.org/p/osf/socarx/ve4z7.html.
[4] R. Chetty, J. Friedman, N. Hendren, and M. Stepner, The economic impacts of COVID-19: Evidence from a new public database built from private sector data, Opportunity Insights (2020), https://www.nber.org/system/files/working_papers/w27431/w27431.pdf.
[5] Curriculum Associates, Understanding student needs, early results from fall assessments, Research Brief, 2020. URL https://tinyurl.com/ybd7ftg9. Accessed December 2020.
[6] J. R. Hoehn and H. J. Lewandowski, Incorporating writing in advanced lab projects: A multiple case-study analysis, Phys. Rev. Phys. Educ. Res. **16**, 020161 (2020).
[7] B. M. Zwickl, N. Finkelstein, and H. J. Lewandowski, The process of transforming an advanced lab course: Goals, curriculum, and assessments, Am. J. Phys. **81**, 63 (2013).
[8] M. F. J. Fox, A. Werth, J. R. Hoehn, and H. J. Lewandowski, Teaching labs during a pandemic: Lessons from Spring 2020 and an outlook for the future, arXiv:2007.01271.
[9] F. R. Bradbury and C. F. J. Pols, A pandemic-resilient open-inquiry physical science lab course which leverages the maker movement, arXiv:2006.06881.
[10] T. Feder, Universities overcome bumps in transition to online teaching, Phys. Today **73**, No. 6, 22 (2020).
[11] P. Klein, L. Ivanjek, M. Nikolay Dahlkemper, K. Jelii, M.-A. Geyer, S. Kchemann, and Ana Susac, Studying physics during the COVID-19 pandemic: Student assessments of learning achievement, perceived effectiveness of online recitations, and online laboratories, arXiv:2010.05622.
[12] S. Shivam and K. Wagoner, How well do remote labs work? A case study at Princeton University, arXiv:2008.04499.
[13] L. Leblond and M. Hicks, Designing laboratories for online instruction using the iOLab Device, Phys. Teach. **59**, 351 (2021).
[14] B. M. Zwickl, T. Hirokawa, N. Finkelstein, and H. J. Lewandowski, Epistemology and expectations survey about experimental physics: Development and initial results, Phys. Rev. ST Phys. Educ. Res. **10**, 010120 (2014).
[15] J. Kozminski, H. J. Lewandowski, N. Beverly, S. Lindaas, D. Deardorff, A. Reagan, R. Dietz, R. Tagg, M. Eblen-Zayas, J. Williams, R. Hobbs, and B. Zwickl, AAPT Recommendations for the Undergraduate Physics Laboratory Curriculum Subcommittee Membership, Technical report (American Association of Physics Teachers Committee on Laboratories, College Park, MD, 2014), https://www.aapt.org/Resources/upload/LabGuidlinesDocument_EBendorsed_nov10.pdf.
[16] E. Etkina, B. Gregorcic, and S. Vokos, Organizing physics teacher professional education around productive habit development: A way to meet reform challenges, Phys. Rev. Phys. Educ. Res. **13**, 010107 (2017).
[17] W. Sandoval, Science education's need for a theory of epistemological development, Sci. Educ. **98**, 383 (2014).
[18] B. Devos, Key Policy Letters Signed by the Education Secretary or Deputy Secretary, March 2020, https://www2.ed.gov/policy/gen/guid/secletter/200320.html.
[19] B. R. Wilcox, B. M. Zwickl, R. D. Hobbs, J. M. Aiken, N. M. Welch, and H. J. Lewandowski, Alternative model for administration and analysis of research-based assessments, Phys. Rev. Phys. Educ. Res. **12**, 010139 (2016).
[20] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.17.010148 which contains three Appendices. Appendix A contains an analysis of student responses to the remote lab survey questions in order to understand the types of lab activities students partook in during 2020. Appendix B presents the results of linear modeling to control for possible predictors of the E-CLASS post-test score (the pretest score, major, and gender) to isolate the effect of courses in 2020 compared to 2019. Appendix C presents the effect size calculations for individual questions.
[21] 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 (2016).
[22] E. F. Redish, J. M. Saul, and R. N. Steinberg, Student expectations in introductory physics, Am. J. Phys. **66**, 212 (1998).
[23] B. R. Wilcox and H. J. Lewandowski, Research-based assessment of students' beliefs about experimental physics: When is gender a factor?, Phys. Rev. Phys. Educ. Res. **12**, 020130 (2016).
[24] B. R. Wilcox and H. J. Lewandowski, Improvement or selection? A longitudinal analysis of students' views about experimental physics in their lab courses, Phys. Rev. Phys. Educ. Res. **13**, 023101 (2017).
[25] H. B. Mann and D. R. Whitney, On a test of whether one of two random variables is stochastically larger than the other, Ann. Math. Stat. **18**, 50 (1947).
[26] E. E. Cureton, Rank-biserial correlation, Psychometrika **21**, 287 (1956).
[27] J. Cohen, The earth is round (p. 05), Am. Psychol. **49**, 997 (1994).
[28] I. Rodriguez, E. Brewe, V. Sawtelle, and L. H. Kramer, Impact of equity models and statistical measures on interpretations of educational reform, Phys. Rev. ST Phys. Educ. Res. **8**, 020103 (2012).
[29] D. S. Kerby, The simple difference formula: An approach to teaching nonparametric correlation, Comprehensive Psychol. **3**, 11 (2014).
[30] R Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2020, https://www.R-project.org/.
[31] S. Mangiafico, rcompanion: Functions to Support Extension Education Program Evaluation, 2020, https://CRAN.R-project.org/package=rcompanion. R package version 2.3.25.
[32] R. E. McGrath and G. J. Meyer, When effect sizes disagree: The case of r and d, Psychol. Methods **11**, 386 (2006).
[33] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed., (Lawrence Erlbaum, Hillsdale, NJ, 1988).

[34] C. O. Fritz, P. E. Morris, and J. J. Richler, Effect size estimates: current use, calculations, and interpretation, J. Exp. Psychol. Gen. **141**, 2 (2012).

[35] B. R. Wilcox and H. J. Lewandowski, students' epistemologies about experimental physics: Validating the Colorado Learning Attitudes About Science Survey for Experimental Physics, Phys. Rev. Phys. Educ. Res. **12**, 010123 (2016).

[36] J. Neyman and E. S. Pearson, On the use and interpretation of certain test criteria for purposes of statistical inference, Biometrika **20A**, 175 (1928).

[37] R. A. Armstrong, When to use the Bonferroni correction, Ophthalmic Physiolog. Opt. **34**, 502 (2014).

[38] Note 1. We recognize that there is some debate on the use of the Holm-Bonferroni correction [37] and, therefore, our use of it here may be simply considered as a way to identify the items that we are most confident in saying there was a difference in the shifts in scores between 2019 and 2020.

[39] Indiana University Center for Postsecondary Research, The Carnegie classification of institutions of higher education (2018), http://carnegieclassifications.iu.edu/.

[40] Federal Register Vol. 62, No. 210 Thursday, October 30, 1997 Notices, https://www.govinfo.gov/content/pkg/FR-1997-10-30/pdf/97-28653.pdf.

[41] S. Kanim and X. C. Cid, Demographics of physics education research, Phys. Rev. Phys. Educ. Res. **16**, 020106 (2020).

[42] E. A. Vogels, A. Perrin, L. Rainie, and M. Anderson, *53% of Americans say the Internet has been Essential during the COVID-19 Outbreak* (Pew Research Center, Washington, DC, 2020), https://www.pewresearch.org/internet/2020/04/30/53-of-americans-say-the-internet-has-been-essential-during-the-covid-19-outbreak/.

[43] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics Bull. **1**, 80 (1945).

[44] URL https://jila.colorado.edu/lewandowski/research/e-class-survey-statements.

[45] N. G. Holmes and C. E. Wieman, Introductory physics labs: We can do better, Phys. Today **71**, 38 (2018).

[46] Note 2. Which, in any case, we do not believe to be a reasonable argument that any lab instructor would agree with.