Students' views about experimental physics in a large-enrollment introductory lab focused on experimental scientific practices

Nidhal Sulaiman,[1,2,3,*] Alexandra Werth,[1,2] and H. J. Lewandowski[1,2]

[1]*JILA, National Institute of Standards and Technology and the University of Colorado, Boulder, Colorado 80309, USA*

[2]*Department of Physics, University of Colorado Boulder, Boulder, Colorado 80309, USA*

[3]*Department of Physics, College of Science, Sultan Qaboos University, P.O. Box 36, AlKhod 123, Oman*

(Received 4 May 2022; revised 9 November 2022; accepted 16 December 2022; published 2 March 2023)

A large-enrollment, introductory physics laboratory course at the University of Colorado Boulder has undergone a recent transformation to help students' develop lab skills and better align students' views and beliefs about experimental physics with those of expert experimental physicists through engagement with authentic scientific practices. We examine the impact of the transformation on women and men in the course and report the effect of this transformation on the students' views and beliefs using the Colorado Learning Attitudes about Science Survey for Experimental Physics (E-CLASS) as a measurement tool. We analyze over 3000 student responses from both before and after the transformation for both women and men on overall E-CLASS scores, as well as item-by-item. The results show statistically significant increase in the overall average E-CLASS score after the transformation as compared to that of the course before the transformation regardless of gender. In addition, item-by-item analysis indicates that there are larger gains in a few E-CLASS items, especially those related to the new course learning goals and some of these items are different for women and men. Our results show that students can have different lab experiences depending on their identity, an important aspect that should to be taken into account when designing educational interventions.

DOI: 10.1103/PhysRevPhysEducRes.19.010116

I. INTRODUCTION

Physics lab courses are vital to undergraduate physics degree programs. These labs can allow students to experience authentic, hands-on physics while, importantly, giving them an opportunity to see what it means to be a physicist and develop students' sense of belonging [1–3], identity [4], and epistemology (i.e., beliefs about the nature of learning and the process of knowing physics) [5–7]. Furthermore, they can provide students a unique opportunity to acquire experimental skills essential for their future careers, such as understanding measurement uncertainty, exploring and troubleshooting experimental apparatus, working in teams, and developing communication skills [8,9].

Given these factors, physics lab courses, particularly at the introductory level, can strongly shape students' views of experimental physics. Considering the impact a course can have on students' epistemologies surrounding experimental physics, designing effective physics labs at the introductory level is important—especially for women and other marginalized groups in STEM—as students' beliefs and expectations about the nature of doing and knowing science (i.e., epistemologies) has been linked to decisions to continue to pursue the sciences (i.e., retention and persistence) [10–13].

Understanding the effects of lab experiences in the physical sciences on different student populations is crucial, as we aim to make our classes more inclusive and equitable. For example, a recent study [14] observed that within the context of an introductory physics lab, women and men assume different roles and, without prior direction, men tend to interact with the lab instruments, while women work on communication-related aspects, such as writing and presentations. Thus, it is paramount to consider the wide variety of ways that differing student populations may engage in a course when designing learning experiences—particularly if one is designing new learning experiences to address inequities based on identity.

The University of Colorado Boulder (CU Boulder) recently embarked on transforming the large-enrollment, introductory lab (PHYS 1140) for engineering and physical science majors. Prior to the transformation, this one-credit

*nidhal.sulaiman@colorado.edu

course focused on teaching students how to propagate errors—a goal that was not deeply valued by either the students or the physics and engineering faculty members [15]. The transformed course, instead, was developed based on achieving five main goals: (i) Students' epistemology of experimental physics should align with expert views, (ii) students should have a positive attitude about the course, (iii) students should have a positive attitude about experimental physics, (iv) students should be able to make a presentation quality graph showing a model and data, and (v) students should demonstrate a setlike reasoning when evaluating measurements [15]. Furthermore, like most universities, CU struggles with the retention of women in STEM, and women represent a minority of the students taking its introductory lab course (roughly 25%–30% of the entire class). Through this work, we study how the course transformation at CU Boulder impacted both women and men—specifically in regards to the first goal of developing more expertlike epistemologies of experimental physics.

Through this work, we answer two research questions by utilizing the Colorado Learning Attitudes of Science Survey for Experimental Physics (E-CLASS) [10], a widely used survey on epistemology and expectations for experimental physics:

RQ1. Did the course transformation impact women's and men's overall views of experimental physics?

RQ2. How did the course transformation impact the views of women and men in the course–particularly the views that are aligned with the course goals and design principles?

In addition to looking at the students overall, we look at women and men student populations separately to assess the impact of the PHYS 1140 transformation on their epistemological beliefs and expectations about experimental physics. This analysis protocol was chosen to explicitly avoid looking for gender gaps or "gap gazing," as it has been called for in the community [16,17]. This perspective can be used to see improvements of a particular group of students' experience and learning, while avoiding the gap gazing pitfalls (Sec. II A). This focus is useful for exploring individual group improvements, while also being able to observe potential inequities in the course.

In this paper, we first provide background looking at gap gazing and its potential pitfalls (Sec. II A) and the context for the introductory physics lab at CU Boulder both before and after the transformation in Sec. II B. Next, we present the demographics of student population in the before transformation (BT) and after transformation (AT) courses, as well as the data collection method in Sec. III B. In Sec. III C, we describe our analysis method, where we present three linear regression models, where results for women and men are presented separately in Secs. V and VI. In Sec. VII we follow the results with speculating on how the transformed course positively increased E-CLASS scores and many of the individual E-CLASS items, as well as why the course transformation may have impacted women and men differently. Finally, we conclude with a future outlook for both lab instructors and researchers in Sec. VIII.

II. BACKGROUND

A. Gap gazing in physics education research

The underrepresentation of women and other marginalized racial and ethnic groups has received growing attention from the physics education research community (PER) [18–22]. The effort to understand the lack of this representation is often focused on the performance gap [23]; for example, by comparing women's prescores in concept inventories [23–25] with those of men. This is referred to as gap gazing. As a result, more attention is placed on closing the performance gap, mostly by taking middle-class white men as the performance standard [16]. These studies tend to reinforce a deficit model of education about under-represented groups [17] and generally do not report student experiences, nor the roles they assume, in their classes and labs. For example, multiple identities (intersectionality) may have visible effects on students' E-CLASS scores [26], such as whether or not a student is a woman *and* a physics major. Therefore, *directly* and *exclusively* comparing women's and men's performance in a class can have misleading findings that do not represent the entire picture.

Furthermore, gap gazing studies generally do not track students throughout their career in physics programs. Nevertheless, recent efforts in this direction have emerged [27], which found that women physics majors are more likely than men to receive a physics degree and are on par with men in passing upper level courses.

Gender differences, even if small, could become more significant through their reinforcement of gender stereotype threats [28,29]. These threats can have profound effects on women's self-efficacy and confidence in STEM courses, negatively impacting their performance [30]. This mirrors the finding that marginalized students are more inclined to have less confidence in STEM fields, even when they are equipped with the necessary skills and abilities [31–33]. Interestingly, the study by Mujtaba and Reiss [33] shows that stereotype threats can also lead to biases, where young women receive less encouragement from teachers, family, and friends to study physics in comparison with men. This also affects teachers' views as to what mechanisms play a role in student success, where boys' "brightness" is attributed to their success in physics, while hard work is seen as the factor propelling girls to do well in the field [34].

It is worthwhile to mention that there is no reason to expect women and men to have the same learning experience in the same educational setting, as they may change their attitudes as per their cultural and societal

influences [9]. Experiences can, of course, also vary within a particular gender [20].

Overall, gap gazing also raises a variety of issues when conducting research, and gender-gap studies must be approached with caution and care. It is important to consider these aspects when interpreting gap gazing research in order to further develop physics curricula and best practices within teaching and learning. One way to reduce the negative aspects of gap gazing while still studying the inequities in curricula for women and other marginalized racial and ethnic groups is by conducting data analysis that does not directly compare populations (i.e., not holding the majority students' as the standard of comparison in the model).

In this work, we conduct three separate regression models relating E-CLASS post-test scores to course type, while accounting for pretest scores. The three models look at (a) the entire class, (b) women, and (c) men, as to not hold men as the standard in the regression model. We then compare only the similarities and differences in the types of items that showed significant changes in the regression models rather than directly comparing numbers and gains. Although this work is fundamentally quantitative in nature and seeks to explain *what* changed from before to after the transformation, we include a discussion and call for more explanatory, qualitative analysis that would help us identify the reasons for these gendered differences and make improvements to the course in the future.

B. Course context

The course studied in this work is a one-credit, introductory, stand-alone lab course at CU Boulder. The enrollment in the course is typically around 500–700 engineering and physical science majors. It is often taken in the same semester as the calculus version of General Physics 2 (second semester of introductory physics) and covers physics topics across both semesters of introductory physics, including mechanics, electricity and magnetism, and optics. The course includes one two-hour lab section each week and six lectures spread throughout the semester. Each lab section has between 16–20 students and one graduate student teaching assistant (TA).

The original course before the transformation (or BT course) had been adapted over many years from one created in the 1960s, which had over 100 experiments for the students to choose from. By 2016, the number of experiments had been reduced to six experiments that the students completed. The experiments mostly required students to reproduce a well-known result, such as the index of refraction of Lucite or how the period of a pendulum changed as a function of length. Because of lack of equipment, students did not complete the experiments in any particular order, but based on a schedule for the class, which had them rotate through the lab experiments. They took the data for the experiment in one week and then spent the lab time the following week doing the analysis and writing a lab report using the *Mathematica* software package. The lab manuals were very prescriptive and outlined the exact analysis the students should do for their report, including which sections should be included. Most of the analysis was on propagating the uncertainty in measurements and the formulas for this were given to the students in the lab manuals. The students worked alone on the experiments and reports, as there was fear from some instructors that students would plagiarize the work of a lab partner if given the chance to work together. In addition to the lab reports, students completed homework assignments on error propagation using the partial differential formalism and statistical analysis, as well as prelab activities related to the experiments they performed in lab. Overall, this version of the course could be considered very traditional and likely typical of many such introductory lab courses across the United States. The course was not liked nor considered useful by students (based on end-of-course feedback) and instructors teaching the course.

An effort to completely transform the course to be better aligned with the needs of the students in departments served by the course began in 2016, with the first semester of the new course being offered in the Spring 2018 semester. Details of the transformation process can be found in previous work [15]. One of the early outcomes of this process was a set of learning goals (as listed in the Introduction) and additional guiding principles, which included items that were explicitly not goals for the course and constraints that were present in our local context (although may be present in other similar institutions). The guiding principles were as follows:

- Students will not write full formal lab reports.
- Students will not choose which experiments they do each week.
- Students will not be required to learn the details of coding in *Mathematica*, Matlab, etc.
- Students will not be required to learn to derive differential propagation of errors formalism.
- Students will be required to put in effort appropriate for a one-credit course, which includes no more than three hours outside of scheduled class time.
- Learning experimental or engineering design are not goals for the course.
- The new structure should recognize the limitations of the skills and time of the TAs.
- The new course should be sustainable in its new format even when many different faculty members rotate teaching the course.
- This course must remain an "experimental experience" to satisfy ABET accreditation.
- The physics topics covered in the lab should include material from both General Physics 1 and 2 lecture courses.

The resulting course after the transformation (AT course) had little in common with the BT course beyond the structure of one two-hour lab session per week and six lectures. The lab sections were still taught by graduate TAs, but also now included an undergraduate learning assistant [35]. The students worked in pairs and collected data together, but wrote in their own electronic lab notebook that they turned in for grading each week.

The course consisted of 12 one-week labs grouped into four sections: skill building, mechanics, electronics, and optics topics [15]. The skill building labs included opportunities for students to learn how to keep an electronic lab notebook (OneNote), fit data to a model and create a graph, and develop proficiency with basic statistical analysis through an experiment that allowed for acquisition of thousands of data points, which were used to calculate the standard deviation and standard error. These skills are all aligned with authentic scientific practices of documentation [36,37] and modeling of experiments [38]. The mechanics labs used projectile motion equipment to explore statistical uncertainty by comparing the predicted spread of measurements to the estimated uncertainty. The electronics labs had students confront systematic errors, as well as some components of modeling [39]. Finally, the optics labs had students make decisions based on measurements and uncertainty, as well as additional modeling components.

The lab guides were still rather prescriptive, but instead of having a step-by-step list of tasks for students to follow, the structure was more aligned with authentic scientific practices. To accomplish this, the lab guides had the following sections: Explore, predict, gather and analyze data, discuss, compare, Draw conclusions, present, and reflect. Not all labs had all components and most labs had multiple instances of some of these sections. An important feature of these guides was the "compare" phase. In no case did students compare a measured value to a "known or correct" value. In many cases, students compared to measurements from other groups in the same section. For example, for the Snell's law lab, students were given a sample of sugar water with an unknown concentration. They were asked to determine the concentration and uncertainty (based on the index of refraction) and record that on a central spreadsheet. Near the end of the lab section, the students were asked to reason about how many different sugar mixtures the class was given based on the measurements (and uncertainties) from all lab groups. Another important component was the "discuss" phase, where students were often asked to find another lab group to discus their procedure, data, or results. They would then have the opportunity to revise their work based on those discussions.

In addition to the lab notebooks that were turned in at the conclusion of the lab session each week, students watched short (∼10 min) prelab videos where one of the initial instructors of the course would demonstrate the equipment and go over any related concepts needed for the lab [40]. There were embedded questions within the videos the students had to answer and these were graded for correctness. Finally, there were six lectures on various topics directly relevant for the labs, such as statistical and systematic error analysis and the physics of fiber optics, as students would not have seen that in their lecture courses. Concept tests with the iClicker technology were used extensively in the lectures, where students were encouraged to work in groups to answer the questions.

III. METHODOLOGY

A. E-CLASS

To measure the impact of the courses on students' views of experimental physics, we used a research-based assessment, E-CLASS [10]. E-CLASS is an epistemology and expectations survey focused on experimental physics labs at the undergraduate level. There have been several studies showing evidence of validity [10,11] and it has been used widely by both in the United States and internationally [41,42].

The survey is administered via an online system at the beginning (pretest) and the end (post-test) of a semester [43,44] to measure the change in students' views. The survey consists of 30 items, such as "*When doing an experiment, I just follow the instructions without thinking about their purpose.*" The students rank their level of agreement with the statement on a five-point Likert scale from "strongly agree" to "strongly disagree" based on two questions: *What do YOU think when doing experiments for class?* and *What would experimental physicists say about their research?* For this study, we use responses to only the first question. Additionally, in the post-test, E-CLASS asks students what was important for earning a good grade in the class and asks students to optionally provide demographic information such as gender, race or ethnicity, and major.

The items are scored using the established expertlike response for that item [10]. The responses for each item are collapsed to a three-point scale, where "strongly (dis)agree" and "(dis)agree" are combined. Points are then given based on agreement with experts: $+1$ point for responses consistent with experts, $+0$ for neutral, and -1 for responses inconsistent with experts. A student's overall E-CLASS score is given by the sum of the scores on each item resulting in a possible range of scores from $[-30, 30]$. We also calculate the average score per question for a group of students, resulting in an average in the range $[-1, 1]$.

The survey was designed such that the statements covered a broad range of learning goals for college-level labs from introductory courses to advanced undergraduate labs. As such, not all statements may be important for a particular lab course. The statements related to the learning goals for the transformed course studied here are listed in

TABLE I. E-CLASS items that are related to learning goals of the AT course.

No.	Item
5	Calculating uncertainties usually helps me understand my results better.
7	I don't enjoy doing physics experiments.
9	When I approach a new piece of lab equipment, I feel confident I can learn how to use it well enough for my purpose.
13	If I try hard enough I can succeed at doing physics experiment.
16	The primary purpose of doing physics experiments is to confirm previously known results.
18	Communicating scientific results to peers is a valuable part of doing physics experiments.
19	Working in a group is an important part of doing physics experiments.
22	If I am communicating results from an experiment, my main goal is to make conclusions based on my data using scientific reasoning.
23	When I am doing an experiment, I try to make predictions to see if my results are reasonable.

Table I [45], and we examine both these statements, as well as others on the E-CLASS for this work.

B. Data collection

We gathered data from three offerings (Fall 2016, Spring 2017, and Fall 2017) of PHYS 1140 before the course transformation, resulting in a total of 2222 students, and from four offerings (Spring 2018, Fall 2018, Spring 2019, and Fall 2019) after the course transformation, giving a total of 1748 students. This is in order to have enough statistical power to enable us to draw conclusions, especially for women who make up about 25% BT and 30% AT of the course registrants. All the BT offerings were taught by the same instructor. This instructor also taught the AT Spring 2018 and Fall 2018 offerings, while another instructor taught the AT Spring 2019 and Fall 2019 offerings. Pretest and post-test responses were matched based on students' names and ID numbers. We use only matched data for our analysis.

The student population demographics, such as gender and ethnicity or race are self-reported from the students at the end of the post-test (to avoid triggering stereotype threat [46,47]). A total of 1449 students (382 women and 1067 men) in the BT offerings and 1723 students (522 women and 1201 men) in the AT offerings completed both the pretest and post-test surveys, and formed the dataset analyzed here. We note that the survey's question on gender type included "other" as a third category. We acknowledge that gender is a spectrum and does not have a binary division; however, because the other category had so few responses (∼1% of all responses), we excluded it here and performed our analysis with the binary gender data to avoid making any nonsignificant conclusions.

The demographic distributions of women and men race or ethnicity and major are also shown in Tables II and III, respectively. About half of the women students in the introductory physics lab course were engineering majors and about 40% were other science majors. Less than 5% of the women enrolled in the course were physics majors.

For men, about 50% were engineering majors and 30% were other science majors. About 10% of the men enrolled in the course were physics majors, as shown in Table III.

TABLE II. Demographic data of the women enrolled in the introductory physics lab course in its BT (Fall 2016 to Fall 2017) and AT (Spring 2018 to Fall 2019) format with $N_{BT} = 382$ and $N_{AT} = 522$.

	BT(%)	AT(%)
American Indian or Alaska native	0.8	0.8
Asian American	19.4	17.4
Black or African American	1.6	1.1
Hispanic/Latino	9.4	7.3
Native Hawaiian or other Pacific Islander	0.8	0.0
White	69.4	73.0
Other race or ethnicity	2.6	2.5
Physics	4.5	4.6
Engineering	54.5	53.8
STEM	37.4	38.9
Other disciplines	2.9	2.7
Undeclared	0.8	0.0

TABLE III. Demographic data of the men enrolled in the introductory physics lab course in its BT (Fall 2016 to Fall 2017) and AT (Spring 2018 to Fall 2019) format with $N_{BT} = 1067$ and $N_{AT} = 1201$.

	BT(%)	AT(%)
American Indian or Alaska native	1.2	0.7
Asian American	15.7	14.3
Black or African American	1.6	1.4
Hispanic/Latino	8.8	8.5
Native Hawaiian or other Pacific Islander	0.7	0.7
White	70.0	74.4
Other race or ethnicity	3.7	2.1
Physics	8.9	10.8
Engineering	55.2	52.0
STEM	31.4	32.6
Other disciplines	3.7	4.3
Undeclared	0.8	0.2

C. Analysis method

Throughout this work, we determine statistically significant differences and effect sizes (i.e., the practical significance) along with the corresponding confidence interval between E-CLASS scores of the BT and AT course. We do this for men and women separately using two linear regression models where BT or AT and pretest scores are input variables. We calculate p values for overall test scores and scores of individual E-CLASS items to determine whether or not there is a statistically significant difference between the AT and BT courses. We discuss the effect size, g of the items with statistically significant differences and the potential cause of these shifts from the BT to AT course. Below, we elaborate on these two analysis tools used to determine the effect of the course transformation on students' views of experimental physics.

1. Linear regression model

We use the linear regression model called ANCOVA (or analysis of covariance) to understand how the E-CLASS scores are affected by different variables, such as being before transformation or after transformation [26]. Linear regression models are commonly used for this type of research (see, for example, Refs. [9,48]). We also report the 95% confidence intervals for the estimated fit parameters, which is helpful in determining if these estimates are significantly different from zero.

Recall that one of our broad research questions is how much the transformation of course impacted the students' views about experimental physics at the end of each of the BT and AT courses for women and men groups separately. To this end, we have applied ANCOVA analysis on women's and men's responses separately. A commonly used approach for looking at gender differences in data is using gender as a categorical variable. However, we choose to run two separate regressions in order to not hold men as the "standard" in the regression model. Still, using this approach, we are able to understand women's and men's responses both separately and through *qualitative* comparison of items that showed significant changes. All of our analysis was done using R (the regression analysis package) [49].

2. Effect sizes

We use Hedge's g (not to be confused with Cohen's d [50]) to measure the magnitude of the effect resulting from a course transformation. The effect size gives the difference of the mean scores of the two groups $(m_A - m_B)$ in terms of their standard deviations. Because the two different groups in this study have different sample sizes, one needs to weight each group's standard deviation by its sample size [51]. So, the Hedge's g defined as

$$g = \frac{(m_A - m_B)}{s_{pg}}, \tag{1}$$

where

$$s_{pg} = \sqrt{\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{(n_A + n_B - 2)}}, \tag{2}$$

where n_A and n_B are the sample sizes of groups A and B, respectively.

The effect size g itself is an estimate, so one also needs to provide the confidence interval quantifying its uncertainty [24,52]. Usually, one takes the 95% confidence interval, corresponding to a significance level $\alpha = 0.05$ (or 5%). This value of α is the probability the estimate lies outside the confidence interval [53]. The 95% confidence interval, $\mathrm{CI}_{95\%}$, is

$$\mathrm{CI}_{(95\%)} = g \pm s_g * 1.96. \tag{3}$$

We, additionally, report the corrected 95% confidence interval using the Ryan-Holm step-down Bonferroni procedure for the item-by-item analysis [54].

The standard error, s_g, for the Hedge's g statistics is calculated using

$$s_g = \sqrt{\frac{(n_A + n_B)}{(n_A n_B)} + \frac{g^2}{2(n_A + n_B)}}. \tag{4}$$

In this equation, the first term under the square root sign reflects the uncertainty in the estimate of the mean difference [the numerator in Eq. (1)], whereas the second term gives the uncertainty in the estimate of s_{pg} [Eq. (2)].

Furthermore, we use the F-Statistic to compare two variances, s_A and s_B, by taking their ratio. In particular, the F-Statistic looks at whether the variance between the means of the BT and AT courses is significantly different. We also choose to report R^2, the coefficient of multiple determination [55]. This is commonly used with multiple regression analysis and it gives the percentage of variation in the predictor variables that is explained by a linear model.

3. Modeling post-test scores while controlling for pretest scores

We also want to isolate the effect of incoming students' views. This is important as student preparation prior to enrolling in lab courses can be different depending on previous educational experience, which affects E_{post} scores. To do this, we use an ANCOVA, which is represented mathematically as

$$E_{\mathrm{post}} = \beta_0 + \beta_1(\mathrm{CourseType}) + \beta_2 E_{\mathrm{pre}} + \epsilon. \tag{5}$$

Here, the CourseType is a categorical variable that we chose to be 0 for the BT course and 1 for the AT course, ϵ is a residual error, and β_1 can be interpreted as effect sizes in standardized units for course type post-test score.

Equation (5) gives the expected average post-test E-CLASS scores for students in the BT and AT courses who have the same pretest scores. In our case, ANCOVA achieves this by splitting the pretest scores into two subgroups based on CourseType, one for BT pretest scores and one for AT pretest scores, and calculating the mean for each CourseType. These means are then adjusted suitably to predict the estimated post-test scores [48].

The fit parameters resulting from the ANCOVA models are unstandardized because they are in the units of E-CLASS scores (out of 30). To allow for comparisons within and among different models, variables, and studies, we also compute the z scores, which normalize (or, standardize) the deviation of a student's E-CLASS score $x^{(i)}_{\text{student}}(i = A, B)$ from the survey average score m_i in units of the standard deviation s_i of the students' scores distribution,

$$z^{(i)}_{\text{student}} = \frac{(x^{(i)}_{\text{student}} - m_i)}{s_i}. \tag{6}$$

These standardized results are presented in Appendices A, B, and C. This standardizing also has the added advantage of reducing gap gazing.

Additionally, we run ANCOVA analysis on individual E-CLASS items. For these individual items, E_{post} and E_{pre} are Likert-scale values given as -1 for disagree, 0 for neutral, and 1 for agree. While treating Likert-scale data as interval is contended [56], especially when looking at single items, many agree that parametric statistics, such as ANCOVA, can be used with minimal concern [56]. However, given that using ANCOVA for single, Likert-scale items is not a standard practice, we have provide additional analysis using Mann Whitney U tests [57] to compare the BT pre scores to the AT pre scores and BT post scores to AT post scores for each of the individual items in Appendix E. Using the Mann Whitney U test, we find—for all practical purposes for which we draw our conclusions—equivalent results to the ANCOVA[1]; however, since there are significant differences in some of the BT pretest scores to the AT pretest scores, we choose to present the ANCOVA results throughout this paper which controls for the pretest scores. This allows us to calculate effect sizes based on the estimated marginal means (also known as covariate-adjusted means) [58].

[1]We also conducted an generalized linear model (GLM) using a logistic regression, but this further collapses what was originally a 5-point Likert-scale onto a binary scale (i.e., comparing "agree" to "neutral" and "disagree"), which reduces our ability to capture the nuance in our data.

The statistical significance of all our results presented in this paper through p values, used the Holm-Bonferroni correction. It is important to note that the Holm-Bonferroni method has a lower increase of type II error risk than the classical Bonferroni method, yet is still a conservative estimator of statistical significance. Given this limitation, in addition to reporting the E-CLASS items that show statistically significant effect sizes and their corresponding p values, we also report the E-CLASS items with statistically nonsignificant effect sizes and their p values as well.

IV. IMPACTS OF THE LAB TRANSFORMATION ON THE ENTIRE CLASS

We report below the overall views of the entire class before and after the transformation to obtain an average picture of the impact the course transformation had on students' epistemological beliefs and expectations of experimental physics. This is then followed by presenting the outcomes for each gender separately.

We apply Eq. (5) to model the post-test E-CLASS scores using a multiple regression. The coefficients of the models are reported in the units of the E-CLASS raw scores, as well as in standardized units, shown in Table VII, Appendix A. Our data meet all the assumptions of multiple regression that allow for unbiased interpretation of statistical significance of the results except for the "homogeneity of variance of residuals" assumption, meaning that the variance in the post-test scores is not constant across all pretest values (heteroskedasticity). Thus, although there will be some bias in our data, we are still able to obtain relevant patterns [59] showing the effect of the transformation. A possible cause for not meeting the heteroskedasticity assumption is that the distributions of the E-CLASS pretest and post-test scores are skewed towards higher scores.

Figure 1 shows the E-CLASS pretest and post-test average scores of the entire class (women and men) for both the BT and AT courses [45], along with the 95% confidence intervals. The overall average BT course yields a statistically significant ($g = -0.11 \pm 0.07$, $p < 0.001$) drop from 18.6 ± 0.2 and 17.9 ± 0.2 ($N = 1449$) in the pretest to post-test. This result clearly indicates a drop in the students' E-CLASS scores suggesting an epistemological shift away from expertlike views in the BT course, where, in contrast, students' views remain the same after the AT course, where the average ($N = 1723$) pre-post scores are 19.0 ± 0.2 and 18.9 ± 0.2 ($p = 0.56$). Even though the goal of the transformation is to align students' views with expertlike views, the transformation has not adversely affected students beliefs, unlike the BT course. This is in line with expectations for attitude surveys; other studies [26] have also shown a similar trend when comparing E-CLASS pretest with post-test scores that are unchanged

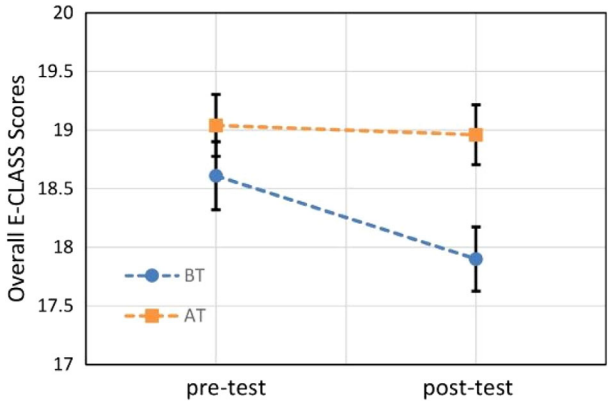

FIG. 1. Comparison of the average overall E-CLASS scores between the BT (blue, $N = 1449$) and AT (orange, $N = 1723$) offerings of the course. Error bars are 95% confidence intervals and dashed lines are for guidance.

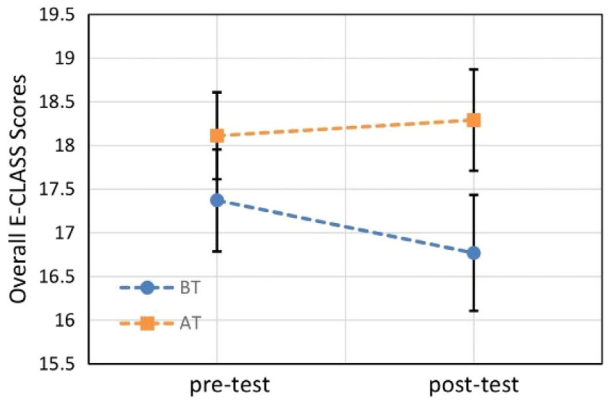

FIG. 2. Average overall pretest and post-test E-CLASS scores for women in the BT (blue, $N = 382$) and AT (orange, $N = 522$) offerings. Error bars are 95% confidence intervals.

when comparing first-year and beyond-first-year college students.

Table IV summarizes the findings of ANCOVA model. The resulting calculation gives a regression slope $\beta_1 = 0.75 \pm 0.18$, which, in standardized form, translates into an effect size of $g = 0.12 \pm 0.03$ ($CI_{95\%} = [0.18, 0.06]$) (also see Table VII, Appendix A). This is statistically significant ($p < 0.001$), meaning the overall class's post-test scores in the AT course are different from overall class's post-test scores in the BT course, holding all other variables constant. Therefore, on average, if two students have similar incoming beliefs they will have different post-test scores, with the student in the transformed course retaining beliefs that are closer to expertlike views, when compared with the BT course.

V. IMPACTS OF THE LAB TRANSFORMATION ON WOMEN'S VIEWS

A. Overall pretest and post-test scores

We start by examining women's performance through their overall pretest and post-test E-CLASS scores in both BT and AT courses independently from men. The raw data are presented in Fig. 2. In order to determine if there is a statistically significant difference between the post-test scores of the BT and AT, we performed an analysis of covariance (ANCOVA) to control for the pretest scores [Eq. (5)]. Table V summarizes the findings of ANCOVA. The resulting calculation gives a regression slope $\beta_1 = 1.0 \pm 0.4$, which, in standardized form, translates into an effect size of $g = 0.15 \pm 0.05$ ($CI_{95\%} = [0.25, 0.05]$) (also see Table VIII, Appendix B). This is statistically significant ($p = 0.005$), meaning women's post-test scores in the AT course are different from women's post-test scores in the BT course, holding all other variables constant. Therefore, on average, if two women have similar incoming beliefs they will have different post-test scores, with the student in the transformed course retaining beliefs that are closer to expertlike views, when compared to the BT course.

We have also checked for the possibility of interaction between the pretest scores and CourseType, but have found it is not significant. Therefore, we have not included an interaction term in Eq. (5).

TABLE IV. Results of linear regression (ANCOVA) for the class's overall post-test scores while controlling for pretest scores.

$$E_{post} = \beta_0 + \beta_1(\text{CourseType}) + \beta_2 E_{pre} + \epsilon$$

Predictors	Raw coefficients	Std. error	F	p value
(Intercept), β_0	4.95	0.34		
CourseType (AT), β_1	0.75	0.18	16.6	<0.001
Pretest score, β_2	0.70	0.02	51785.4	<0.001
Residual standard error	5.208			
Adjusted $R^2(\%)$	36			

TABLE V. Results of linear regression (ANCOVA) for women's post-test scores while controlling for pretest scores.

$$E_{post} = \beta_0 + \beta_1(\text{CourseType}) + \beta_2 E_{pre} + \epsilon$$

Predictors	Raw coefficients	Std. error	F	p value
(Intercept), β_0	4.51	0.59		
CourseType (AT), β_1	1.00	0.35	18.1	0.005
Pretest score, β_2	0.71	0.03	537.8	<0.001
Residual standard error	5.307			
Adjusted $R^2(\%)$	38			

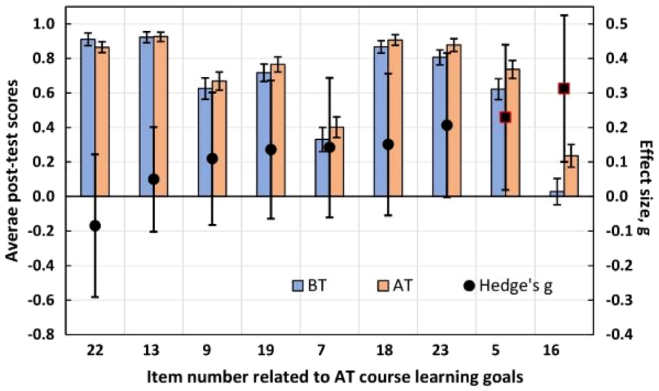

FIG. 3. Women's average post-test scores in the BT (blue) and AT (orange) courses on E-CLASS items (bar graph) that are related to the AT course learning goals. Effect sizes, Hedge's g and the corresponding 95% confidence intervals, are plotted as symbols for each of the items. Items with significant differences between BT and AT post-test scores for women are denoted by squares with red border.

B. Individual E-CLASS items that are related to the AT course learning goals

In addition to the overall score, there are nine E-CLASS items that were identified as related to the AT course leaning goals [60], see Table I. We investigate women's scores for each of these items, as such an understanding may guide further instructional improvements. We conduct an ANCOVA, Eq. (5), for each of the nine items and control for students' pretest scores for that item.

Figure 3 shows post-test average scores for these nine items in the BT and AT courses (bar graph) together with effect sizes, g (represented by the symbols in the figure), and we denote the statistically significant items by squares. These values are also presented in Appendix D, Table X.

Item 16, the only statistically significant change, has a medium effect size of $g = 0.27 \pm 0.21$ reflecting a positive shift towards expertlike views for women in the AT course. This is an important shift, as other studies [61] have found that labs that stress confirmation of results could lead students to "questionable research practices," such as engaging in subjective data interpretation. Additionally, we note that items 5 and 23 have nonzero effect sizes, but the changes are not statistically significant. Further discussion can be found in Sec. VII.

C. Responses to individual E-CLASS items that are not related to the AT course learning goals

It is interesting to look at women's responses to other individual E-CLASS items that are not related to the course learning goals to gain a deeper insight into how women's views have changed in the AT course compared to the BT course more broadly.

As in the previous subsection, we conduct an ANCOVA, Eq. (5), for each one of the remaining 21 E-CLASS items, where we also control for students' pretest scores for that item. We present results for these items in Fig. 4 and Appendix D, Table X.

Only items 17 and 29 have a statistically significant change from BT to AT with medium effect sizes of 0.29 ± 0.21 and 0.21 ± 0.21, respectively. Discussion of factors that may have influenced all of these E-CLASS items can be found in Sec. VII.

VI. IMPACTS OF THE LAB TRANSFORMATION ON MEN'S VIEWS

A. Overall pretest and post-test scores

Overall, men perform similarly to the performance of the entire class shown in Fig. 1, which is not surprising as

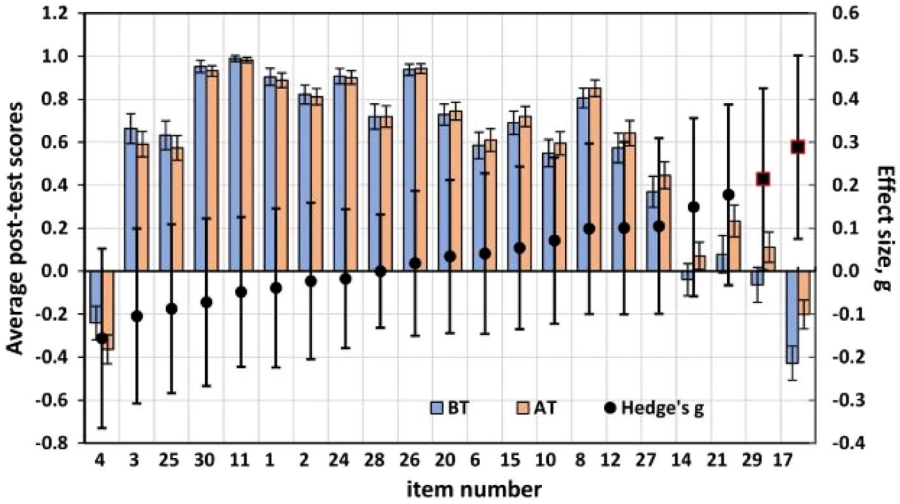

FIG. 4. Women's average post-test scores in the BT (blue) and AT (orange) courses on E-CLASS items (bar graph) that are not related to the AT course learning goals. Effect sizes, Hedge's g and the corresponding 95% confidence intervals, are plotted as symbols for each of the items. Items with significant differences between BT and AT post-test scores for women are denoted by squares.

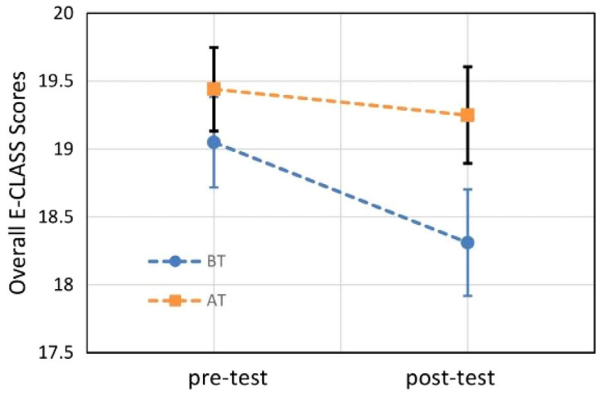

FIG. 5. Average overall pretest and post-test E-CLASS scores for men in the BT (blue, $N = 1067$) and AT (orange, $N = 1201$) offerings. Error bars are 95% confidence intervals.

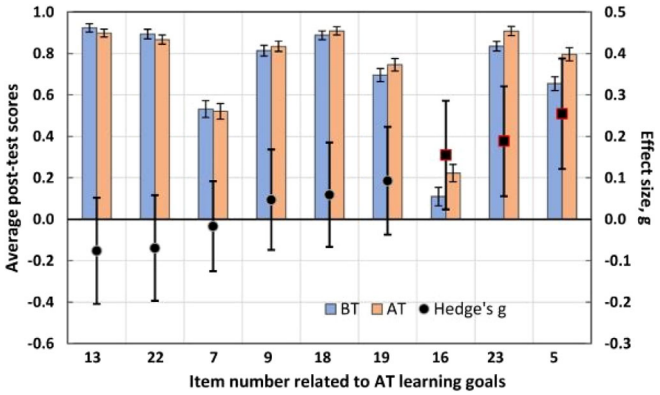

FIG. 6. Men's average post-test scores in the BT (blue) and AT (orange) courses on E-CLASS items (bar graph) that are related to the AT course learning goals. Effect sizes, Hedge's g and the corresponding 95% confidence intervals, are plotted as symbols for each of the items. Items with significant differences between BT and AT post-test scores for men are denoted by square symbol.

they constitute around 75% of the students in the course. Men's raw pretest and post-test scores are plotted in Fig. 5. We perform an ANCOVA [Eq. (5)] to control the incoming beliefs, through pretest scores, to calculate the expected effect of the transformation on E-CLASS scores of men. The results of this analysis are in Table VI. We obtain a regression slope $\beta_1 = 0.67 \pm 0.22$, which in standardized form translates to a Hedge's effect size of $g = 0.10 \pm 0.03$ ($p = 0.002$) (see Table IX, Appendix C for standardized results). The p value is statistically significant, so the men's post-test scores in the AT course are different from the men's post-test scores in the BT course, holding all other variables constant. Additionally, the calculated g has a small, nonzero, effect size indicating that men with similar incoming beliefs will, on average, have slightly higher post-test scores in the AT course than the BT course.

B. Responses to individual E-CLASS items that are related to the AT course learning goals

We examine the men's performance in the nine E-CLASS items related to the AT course learning goals (Table I). As before, we control for incoming men's views by using ANCOVA [Eq. (5)] for these nine items. The results are shown in Fig. 6, where we plot the BT and AT post-test scores, the effect sizes, g, and denote statistically significant differences. Each of these items are also presented in Appendix D, Table XI.

Items 5, 16, and 23 saw significant, positive increases after the implementation of the AT course. We see that item 5 has the largest significant effect size $g = 0.26 \pm 0.13$ among the nine items, as shown in Fig. 6. Items 16 and 23 have a medium effect with $g = 0.16 \pm 0.13$ and $g = 0.19 \pm 0.13$, respectively. However, although item 16 has a significant gain, the average men's post-test AT score remains quite low (nonexpertlike).

TABLE VI. Results of linear regression (ANCOVA) for men's post-test scores while controlling for pretest scores.

$$E_{\text{post}} = \beta_0 + \beta_1(\text{CourseType}) + \beta_2 E_{\text{pre}} + \epsilon$$

Predictors	Raw coefficients	Std. error	F	p value
(Intercept), β_0	5.18	0.41		
CourseType (AT), β_1	0.67	0.22	18.7	0.002
Pretest score, β_2	0.69	0.02	1209.7	<0.001
Residual standard error	5.307			
Adjusted $R^2(\%)$	38			

C. Responses to individual E-CLASS items that are not related to the AT course learning goals

It is interesting to look at men's responses to individual E-CLASS items to further understand how men's views have changed in the AT course as compared to the BT course. Alongside the same analysis we have done for women, here we attempt to focus on items that show significant difference between the two courses.

We find that there are three other E-CLASS items where the differences between the BT and AT post-test scores are statistically significant, see Fig. 7. Items 6 and 14 show small positive, statistically significant gains in favor of the AT course.

VII. DISCUSSION

Our goal of this work is to answer two research questions, **RQ1** "Did the course transformation impact women's and men's overall view of experimental physics?" and **RQ2** "How did the course transformation impact the views of women and men in the course–particularly the views that are aligned with the course

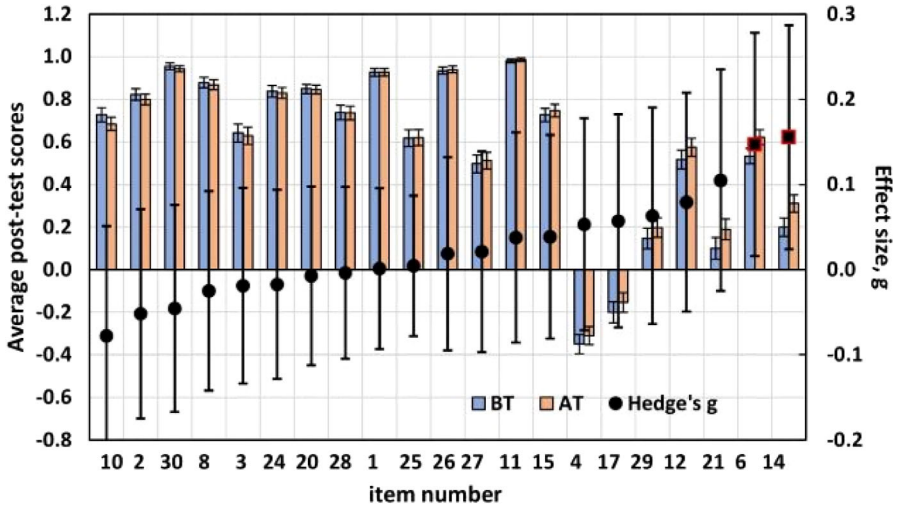

FIG. 7. Men's average post-test scores in the BT (blue) and AT (orange) courses on E-CLASS items (bar graph) that are not related to the AT course learning goals. Effect sizes, Hedge's g and the corresponding 95% confidence intervals, are plotted as symbols for each of the items. Items with significant differences between BT and AT post-test scores for men are denoted by squares.

goals and design principles?" by utilizing the results from the E-CLASS [10].

Through our work we found that

1. The course transformation *did* positively affect both women's and men's overall views surrounding experimental physics.
2. Overall women's and men's views surrounding experimental physics increased equivalently as a result of the course transformation with some individual items increasing significantly more or less for women and men.
3. Both women and men made significant gains in the AT course in regards to the course goals of understanding that the "primary purpose of doing physics experiments is to confirm previously known results" (item 16). Women and men also both had positive, nonzero effects for "calculating uncertainties usually helps me understand my results better" (item 5), and "when I am doing an experiment, I try to make predictions to see if my results are reasonable" (item 23). However, these two items were only statistically significant for men, likely because there were many more men ($N_{BT} = 1067$ and $N_{AT} = 1201$) in the courses than women ($N_{BT} = 382$ and $N_{AT} = 522$).
4. There were significant changes in E-CLASS items, both directly related to the course learning goals and indirectly related to the transformation, which impacted women's and men's views differently.

However, we still have remaining questions about *how* the transformed course positively increased E-CLASS scores in many of the individual E-CLASS items, as well as *why* the course transformation may have impacted women and men differently.

A. How did the transformed course positively increase E-CLASS scores in many of the individual E-CLASS items?

To first understand how the transformed course positively increased E-CLASS scores, we must look at the items that showed significant positive gains for the AT course for either women or men. These items include statements 5, 6, 14, 16, 17, 23, and 29. Of these, items 5, 16, and 23 are related to the course learning goals. In addition, it is important to point out that items 7, 9, 13, 18, 19, and 22 were also related to the course learning goals, but had no significant change for either men or women.

1. Items related to course learning goals that saw significant positive changes after the transformation

From these items related to course learning goals that saw significant positive changes after the transformation (5, 16, and 23), we see that students made gains primarily in items that were specific about understanding measurement uncertainty and how data analysis can impact scientific reasoning and making predictions. This falls in line with our goals and expectations for the course, as calculating uncertainties is an integral part in the transformed AT course, where uncertainties in measured values are stressed in order to *make predictions*.

As an example, for one experiment, students were required to fire a metal ball through several layers of tissue. They launch the ball five times so each time it goes through a single tissue to measure the energy required to rupture it and calculate the standard deviation of the measurements. This is then used to predict the maximum tissue layers needed to stop the ball from penetrating the tissues. This type of experiment could have impacted item 5, "calculating uncertainties usually helps me understand my results better"

because students needed to regularly make *decisions* based on uncertainty, not just report it. Prior research [62] has indicated that students tend to view uncertainties as reflections of imperfections and mistakes done by the experimenter and not as a means to make refinements to obtain better results. However, our results also suggest that it is possible to show students that measurement uncertainties are more than that by asking them to make predictions based on the obtained experimental uncertainties. In the AT course, each group of students is asked to compare its results with those of other groups. At times, the comparison is made throughout the whole lab class. Such comparisons provide a form of experimental repeatability that show students the significance of uncertainties. This is also aided by prompting students to decide on whether to use the standard error on the mean or the standard deviation in many instances in the course.

In addition, the activities in the AT course do not emphasize the verification of known results, as in the BT course. For example, instead of finding the index of refraction of Lucite (which can be simply looked up on the internet), various student groups in the Snell's law lab are given different sugar concentrations in water. Their job is to determine this concentration and to compare with other groups to find which other groups have the same sugar concentration, taking into account the relevant measurement uncertainties. We see this reflected in the positive shift of item 16, which is an important change for the students, as other studies [61] have found that labs that stress confirmation of results could lead students to "questionable research practices," such as engaging in subjective data interpretation and overestimating the size of error bars to have better agreement with theoretical models.

Similarly, unlike the BT course, the lab activities of the AT course put more emphasis on making predictions for the experiments that may have positively impacted item 23, "when I am doing an experiment, I try to make predictions to see if my results are reasonable." All of the labs in the AT course had a "predict" phase. For example, students in the projectiles lab are required to measure a ball's velocity from a launcher and use that measurement to predict the ball's landing position and the associated range corresponding to the launch velocity uncertainty.

2. Items related to course learning goals that did not have significant positive changes after the transformation

Although our intention is to answer how the transformed course *positively* increase E-CLASS scores it would be negligent to not discuss the items related to the goals of the AT course that did not see significant positive change, as these may inform future modifications for both the PHYS 1140 and other labs. The items related to the goal of the AT course that did not have significant positive change were items

7 I don't enjoy doing physics experiments.
9 When I approach a new piece of lab equipment, I feel confident I can learn how to use it well enough for my purpose.
13 If I try hard enough I can succeed at doing physics experiments.
18 Communicating scientific results to peers is a valuable part of doing physics experiments.
19 Working in a group is an important part of doing physics experiments.
22 If I am communicating results from an experiment, my main goal is to make conclusions based on my data using scientific reasoning.

In contrast with the items where we saw gains in the AT course, these items were more related to affect goals (items 7, 9, and 13) and scientific communication (items 18, 19, 22). For example, one of the AT course goals was for students to attain a higher level of enjoyment in these labs (item 7) so as to develop more motivation and positive attitudes [63]; however, it is not clear that this was achieved according to the change in scores of the E-CLASS items. The low and unchanged scores could reflect the fact that the majority of students taking this course were not physics majors and, therefore, their motivation level may have played a role [64]. Understanding how to increase motivation levels requires further studies and a first step could be surveying students as to what activities made the labs more interesting and motivational to them.

3. Items not directly related to course learning goals that saw significant positive changes after the transformation

There were additional items that were not directly related to the course goals, but had significant positive changes between the BT and AT course (items 6, 14, 17, and 29).

For example, the increase in E-CLASS scores for items 17 and 29 that ask about "when I encounter difficulties in the lab, my first step is to ask an expert, like the instructor" and "if I do not have directions for analyzing data, I am not sure how to choose an appropriate analysis method," respectively, both address seeking help when unsure with next steps. Unlike the BT course, the labs in the AT course are designed with "check-in" points so that a student group can check with one or two other groups on procedure and interpretation of results. This actively encourages students to work with each other so that they do not have to rely on an "authority" figure.

In addition, the AT course often prompted students to refer to the equations related to the experiments, which were provided in the lab write-up appendices and covered in the prelab videos. Items 6 and 14 ask, respectively, "scientific journal articles are helpful for answering my own questions and designing experiments" and "when doing an experiment I usually think up my own questions to investigate." Even though our AT course has offered

students some chances to make decisions (e.g., deciding on a best analysis approach to their data analysis), it did not ask students to read journals (item 6) or think of their own questions to investigate (item 14). It is interesting to note that, as a by-product of the AT course implementation, more students affirmed that they formulate their own questions during an experimental investigation. This may be possibly a result of students in the AT course discussing their procedures, analyses, and data interpretations with other groups and then being encouraged to revise based on those discussions. However, more investigation of this point is needed to fully understand why students score higher on these two items in the AT course.

B. Why did the course transformation impact women and men differently?

Throughout our research, it is critical to remember that students do not arrive to a course as an empty vessel; prior research suggests that students' identities and past experiences affect their educational experiences and practices in a physics lab [65]. Thus, it is paramount to consider the wide variety of ways that differing student populations may engage in a course when designing learning experiences. Understanding the effects of lab experiences in the physical sciences on different student populations is crucial, as we aim to make our classes more inclusive and equitable. For this reason, we presented the impact the course transformation had on women and men separately in the results and only compare qualitative differences between the results rather than directly comparing numbers and gains. This perspective can be used to see improvements of a particular group of students' experience and learning, while avoiding the gap gazing pitfalls (Sec. II A).

1. Similarities between women's and men's score changes

Both men and women had significant positive changes in item 16, "the primary purpose of doing physics experiments is to confirm previously known results," having an effect size $g = 0.27 \pm 0.21$ for women and $g = 0.16 \pm 0.13$ for men. It is important to note that items 5, 14, 21, and 23 all have effect sizes of $g > 0.1$ for both men and women (Appendix D). However, only items 5 and 23 are significant for men, while none of the items are significant for women. We postulate that the some of these effects are likely true, particularly for women who only make up 30% of our population, but we do not have the same statistical power to make confident conclusions about these items.

2. Differences between women's and men's score changes

We found "unintended consequences" of transformation, where different lab experiences between the two genders further manifested itself in some E-CLASS items not related to the course learning goals.

6 Scientific journal articles are helpful for answering my own questions and designing experiments.

17 When I encounter difficulties in the lab, my first step is to ask an expert, like the instructor.

29 If I do not have directions for analyzing data, I am not sure how to choose an appropriate analysis method.

One commonality within these three items is their relation to student confidence and agency—aspects of physics labs which have been shown to have a gendered difference [14,20,66,67]. Some studies found that women are less confident of their abilities and are more inclined to take less active roles in experiments, which may have been exacerbated by the fact that women were a minority [66] (in the AT and BT courses). Nevertheless, students can develop more confidence when they are presented with more chances to make decisions during labs—as they likely did in the AT course for items 17 and 29. Unlike the BT course, the AT course had significantly more opportunities for decision making about analysis methods and encouraged students to discuss their procedures and choices within their group and with other lab groups instead of immediately turning to the lab instructor. However, even with the gains in these confidence-related items for women, student views remained far from expertlike and we did not see the same types of gains for women as we did for men in item 6.

We also note, that while not significant item 4, "if I am communicating results from an experiment, my main goal is to have the correct sections and formatting," had a small negative for women and item 19, "working in a group is an important part of doing physics experiments." had a small positive effect for men. If these are in fact true changes between the AT and BT course they would add to our conclusion that perhaps gendered team roles in physics lab courses [14] effect students' views of experimental physics.

VIII. IMPLICATIONS FOR FUTURE RESEARCH

Our results show that the lab experiences of women and men had some similarities, nevertheless, their experiences were not identical despite sharing a common educational setting. In other words, a lab transformation or educational intervention can impact various student populations differently—this result is a crucial finding because it informs of the importance of accounting for the different lab experiences that each gender (or more broadly, various student populations) may have when designing course transformations.

Yet, more work is needed to understand *how* the experiences of women and men differed in the transformed course and *why* this may have lead to different changes in the E-CLASS scores. For example, women made significant gains from BT to AT along E-CLASS items related to confidence. This positive trend needs further study to determine the factors and educational interventions that affected women in this regard in order to enhance their lab experiences in any future course development [68].

Now that we have results classifying the ways in which women and men respond differently to some individual

E-CLASS items, we can use those data to inform directed qualitative studies. The goal of these studies would be to understand possible course structures and activities that are impacting women and men differently. We suggest follow-up studies that include classroom observations, student interviews, and student focus groups, with particular attention to the ways in which women and men approach obstacles and uncertainty in the course.

ACKNOWLEDGMENTS

N. S. would like to thank H. J. L. for her kind hospitality at CU where this work has been done and would also like to thank Sultan Qaboos University for granting a sabbatical leave support at CU. This work was supported by STROBE National Science Foundation Science Technology Center, Grant No. DMR-1548924 and PFC, Grant No. PHY-1734006.

APPENDIX A: ENTIRE CLASS: LINEAR MODEL WITH STANDARDIZED SCORES

TABLE VII. Coefficients for ANCOVA for post-test scores while controlling for pretest scores for all students in the class. The table lists the standardized coefficients and the p values of the ANCOVA results for the CourseType category.

$$E_{\text{post}} = \beta_0 + \beta_1(\text{CourseType}) + \beta_2 E_{\text{pre}} + \epsilon$$

Predictors	Standardized coefficients	Std. error	p value
(Intercept), β_0	−0.06	0.02	0.003
CourseType (AT), β_1	0.12	0.03	<0.001
Pretest score, β_2	0.60	0.01	<0.001
Residual standard error	0.7974		

APPENDIX B: WOMEN: LINEAR MODEL WITH STANDARDIZED SCORES

TABLE VIII. Coefficients for ANCOVA for post-test scores while controlling for pretest scores for women. The table lists the standardized coefficients and the p values of the ANCOVA results for the CourseType category.

$$E_{\text{post}} = \beta_0 + \beta_1(\text{CourseType}) + \beta_2 E_{\text{pre}} + \epsilon$$

Predictors	Standardized coefficients	Std. error	p value
(Intercept), β_0	−0.09	0.04	0.03
CourseType (AT), β_1	0.15	0.05	0.005
Pretest score, β_2	0.61	0.03	<0.001
Residual standard error	0.7873		

APPENDIX C: MEN: LINEAR MODEL WITH STANDARDIZED SCORES

TABLE IX. Coefficients for ANCOVA for post-test scores while controlling for pretest scores for men. The table lists the standardized coefficients and the p values of the ANCOVA results for the CourseType category.

$$E_{\text{post}} = \beta_0 + \beta_1(\text{CourseType}) + \beta_2 E_{\text{pre}} + \epsilon$$

Predictors	Standardized coefficients	Std. error	p value
(Intercept), β_0	−0.06	0.02	0.02
CourseType (AT), β_1	0.10	0.03	0.002
Pretest score, β_2	0.59	0.02	<0.001
Residual standard error	0.8056		

APPENDIX D: ANCOVA RESULTS FOR INDIVIDUAL E-CLASS ITEMS

TABLE X. E-CLASS items compared between BT and AT scores for women as plotted in Figs. 3 and 4. The p values are adjusted using the Holm-Bonferonni correction and the $CI'_{95\%}$ on the effect size are adjusted using the Ryan-Holm step-down Bonferroni procedure. Significance of p values are given by (*) for $p \leq 0.05$, (**) for $p \leq 0.01$, (***) for $p \leq 0.001$, and (****) for $p \leq 0.0001$. The † indicates a small, nonzero effect size ($g < 0.2$) while ‡ denotes a medium, nonzero effect size ($0.2 \geq g < 0.8$). There are no large effect sizes from this analysis.

		Post-test								
		Mean score		Hypothesis testing			Effect size			
#	Item	BT	AT	p value	p value'	g	$CI_{95\%}$		$CI'_{95\%}$	
1	When doing an experiment, I try to understand how the experimental setup works.	0.904	0.888	0.625	1.000	−0.039	−0.171	0.093	−0.224	0.146
2	If I wanted to, I think I could be good at doing research.	0.822	0.811	0.635	1.000	−0.023	−0.155	0.109	−0.205	0.159
3	When doing a physics experiment, I don't think much about sources of systematic error.	0.663	0.591	0.123	1.000	−0.104	−0.237	0.028	−0.308	0.099
4	If I am communicating results from an experiment, my main goal is to have the correct sections and formatting.	−0.241	−0.363	0.025 (*)	0.603	−0.156	−0.289	−0.024†	−0.365	0.052
5	Calculating uncertainties usually helps me understand my results better.	0.622	0.736	0.006 (**)	0.154	0.189	0.057	0.322†	−0.021	0.400
6	Scientific journal articles are helpful for answering my own questions and designing experiments	0.585	0.610	0.531	1.000	0.041	−0.091	0.173	−0.146	0.228
7	I don't enjoy doing physics experiments.	0.330	0.402	0.124	1.000	0.102	−0.030	0.234	−0.100	0.304
8	When doing an experiment, I try to understand the relevant equations.	0.806	0.851	0.160	1.000	0.099	−0.033	0.231	−0.100	0.297
9	When I approach a new piece of lab equipment, I feel confident I can learn how to use it well enough for my purposes.	0.626	0.669	0.302	1.000	0.070	−0.063	0.202	−0.122	0.261
10	Whenever I use a new measurement tool, I try to understand its performance limitations.	0.550	0.594	0.302	1.000	0.071	−0.061	0.203	−0.123	0.264
11	Computers are helpful for plotting and analyzing data.	0.990	0.983	0.721	1.000	−0.048	−0.180	0.084	−0.222	0.126
12	I don't need to understand how the measurement tools and sensors work in order to carry out an experiment.	0.573	0.642	0.142	1.000	0.100	−0.032	0.233	−0.100	0.301
13	If I try hard enough I can succeed at doing physics experiments.	0.923	0.926	0.939	1.000	0.010	−0.122	0.142	−0.142	0.161
14	When doing an experiment I usually think up my own questions to investigate.	−0.040	0.071	0.032 (*)	0.745	0.149	0.017	0.281†	−0.058	0.356
15	Designing and building things is an important part of doing physics experiments.	0.690	0.719	0.441	1.000	0.054	−0.078	0.186	−0.135	0.244
16	The primary purpose of doing a physics experiment is to confirm previously known results.	0.028	0.235	0.000 (****)	0.002 (**)	0.273	0.140	0.406‡	0.060	0.485‡

(Table continued)

TABLE X. *(Continued)*

#	Item	Post-test Mean score BT	Post-test Mean score AT	Hypothesis testing p value	Hypothesis testing p value'	Hypothesis testing g	Effect size $CI_{95\%}$		Effect size $CI'_{95\%}$	
17	When I encounter difficulties in the lab, my first step is to ask an expert, like the instructor.	−0.428	−0.201	0.006 (**)	0.001 (**)	0.289	0.156	0.421‡	0.075	0.502‡
18	Communicating scientific results to peers is a valuable part of doing physics experiments.	0.867	0.907	0.081 (*)	1.000	0.111	−0.021	0.243	−0.094	0.316
19	Working in a group is an important part of doing physics experiments.	0.717	0.766	0.147	1.000	0.096	−0.036	0.228	−0.104	0.296
20	I enjoy building things and working with my hands.	0.728	0.745	0.651	1.000	0.034	−0.098	0.166	−0.144	0.212
21	I am usually able to complete an experiment without understanding the equations and physics ideas that describe the system I am investigating.	0.079	0.233	0.008 (**)	0.213	0.178	0.045	0.310†	−0.032	0.388
22	If I am communicating results from an experiment, my main goal is to make conclusions based on my data using scientific reasoning.	0.911	0.865	0.060 (*)	1.000	−0.125	−0.257	0.008	−0.331	0.082
23	When I am doing an experiment, I try to make predictions to see if my results are reasonable.	0.806	0.878	0.013 (*)	0.334	0.166	0.034	0.299†	−0.043	0.375
24	Nearly all students are capable of doing a physics experiment if they work at it.	0.907	0.901	0.907	1.000	−0.018	−0.150	0.114	−0.179	0.144
25	A common approach for fixing a problem with an experiment is to randomly change things until the problem goes away.	0.632	0.574	0.194	1.000	−0.088	−0.220	0.045	−0.284	0.109
26	It is helpful to understand the assumptions that go into making predictions.	0.938	0.942	0.786	1.000	0.018	−0.114	0.150	−0.150	0.187
27	When doing an experiment, I just follow the instructions without thinking about their purpose.	0.370	0.446	0.120	1.000	0.105	−0.027	0.237	−0.099	0.309
28	I do not expect doing an experiment to help my understanding of physics.	0.719	0.719	0.999	1.000	0.000	−0.132	0.132	−0.132	0.132
29	If I don't have clear directions for analyzing data, I am not sure how to choose an appropriate analysis method.	−0.063	0.112	0.002 (**)	0.044 (*)	0.214	0.081	0.346‡	0.002	0.425‡
30	Physics experiments contribute to the growth of scientific knowledge	0.952	0.932	0.281	1.000	−0.072	−0.204	0.060	−0.267	0.123

TABLE XI. E-CLASS items compared between BT and AT scores for men as plotted in Figs. 6 and 7. The p values are adjusted using the Holm-Bonferonni correction and the $CI'_{95\%}$ on the effect size are adjusted using the Ryan-Holm step-down Bonferroni procedure. Significance of p values are given by (*) for $p \leq 0.05$, (**) for $p \leq 0.01$, (***) for $p \leq 0.001$, and (****) for $p \leq 0.0001$. The † indicates a small, nonzero effect size ($g < 0.2$), while ‡ denotes a medium, nonzero effect size ($0.2 \geq g < 0.8$). There are no large effect sizes from this analysis.

		Post-test								
		Mean	score	Hypothesis testing			Effect size			
#	Item	BT	AT	p value	p value'	g	$CI_{95\%}$		$CI'_{95\%}$	
1	When doing an experiment, I try to understand how the experimental setup works.	0.926	0.927	0.984	1.000	0.001	−0.081	0.084	−0.093	0.096
2	If I wanted to, I think I could be good at doing research.	0.822	0.799	0.224	1.000	−0.052	−0.134	0.031	−0.175	0.071
3	When doing a physics experiment, I don't think much about sources of systematic error.	0.642	0.628	0.628	1.000	−0.019	−0.101	0.064	−0.134	0.096
4	If I am communicating results from an experiment, my main goal is to have the correct sections and formatting.	−0.351	−0.310	0.191	1.000	0.053	−0.029	0.136	−0.071	0.178
5	Calculating uncertainties usually helps me understand my results better.	0.655	0.796	0.000 (****)	0.000 (****)	0.255	0.172	0.337‡	0.122	0.388‡
6	Scientific journal articles are helpful for answering my own questions and designing experiments	0.534	0.622	0.000 (***)	0.011 (*)	0.147	0.064	0.229†	0.016	0.278†
7	I don't enjoy doing physics experiments.	0.532	0.520	0.763	1.000	−0.017	−0.099	0.066	−0.125	0.092
8	When doing an experiment, I try to understand the relevant equations.	0.878	0.868	0.620	1.000	−0.025	−0.107	0.057	−0.142	0.092
9	When I approach a new piece of lab equipment, I feel confident I can learn how to use it well enough for my purposes.	0.814	0.834	0.301	1.000	0.047	−0.035	0.130	−0.074	0.168
10	Whenever I use a new measurement tool, I try to understand its performance limitations.	0.727	0.684	0.059	1.000	−0.078	−0.160	0.005	−0.207	0.051
11	Computers are helpful for plotting and analyzing data.	0.980	0.986	0.201	1.000	0.038	−0.045	0.120	−0.086	0.161
12	I don't need to understand how the measurement tools and sensors work in order to carry out an experiment.	0.517	0.575	0.062	1.000	0.079	−0.003	0.162	−0.049	0.208
13	If I try hard enough I can succeed at doing physics experiments.	0.924	0.898	0.066	1.000	−0.076	−0.159	0.006	−0.204	0.052
14	When doing an experiment I usually think up my own questions to investigate.	0.199	0.311	0.000 (***)	0.006 (**)	0.156	0.073	0.238†	0.024	0.287†
15	Designing and building things is an important part of doing physics experiments.	0.726	0.746	0.390	1.000	0.039	−0.044	0.121	−0.081	0.158
16	The primary purpose of doing a physics experiment is to confirm previously known results.	0.109	0.223	0.000 (***)	0.009 (**)	0.155	0.072	0.237†	0.024	0.286†
17	When I encounter difficulties in the lab, my first step is to ask an expert, like the instructor.	−0.201	−0.154	0.176	1.000	0.057	−0.025	0.140	−0.068	0.182
18	Communicating scientific results to peers is a valuable part of doing physics experiments.	0.888	0.909	0.152	1.000	0.059	−0.023	0.142	−0.067	0.185
19	Working in a group is an important part of doing physics experiments.	0.696	0.746	0.030 (*)	0.708	0.093	0.010	0.175†	−0.037	0.223
20	I enjoy building things and working with my hands.	0.848	0.846	0.890	1.000	−0.007	−0.090	0.075	−0.112	0.098

(Table continued)

TABLE XI. *(Continued)*

#	Item	Post-test Mean score BT	Post-test Mean score AT	Hypothesis testing p value	Hypothesis testing p value'	g	Effect size $CI_{95\%}$		Effect size $CI'_{95\%}$	
21	I am usually able to complete an experiment without understanding the equations and physics ideas that describe the system I am investigating.	0.099	0.189	0.013 (*)	0.330	0.105	0.022	0.187‡	−0.025	0.235
22	If I am communicating results from an experiment, my main goal is to make conclusions based on my data using scientific reasoning.	0.894	0.867	0.096	1.000	−0.069	−0.152	0.013	−0.197	0.058
23	When I am doing an experiment, I try to make predictions to see if my results are reasonable.	0.835	0.908	0.000 (***)	0.000 (***)	0.188	0.106	0.271†	0.056	0.321†
24	Nearly all students are capable of doing a physics experiment if they work at it.	0.837	0.829	0.682	1.000	−0.017	−0.100	0.065	−0.128	0.094
25	A common approach for fixing a problem with an experiment is to randomly change things until the problem goes away.	0.617	0.620	0.999	1.000	0.004	−0.078	0.087	−0.078	0.087
26	It is helpful to understand the assumptions that go into making predictions.	0.934	0.940	0.635	1.000	0.019	−0.064	0.101	−0.095	0.132
27	When doing an experiment, I just follow the instructions without thinking about their purpose.	0.497	0.512	0.617	1.000	0.021	−0.061	0.104	−0.097	0.139
28	I do not expect doing an experiment to help my understanding of physics.	0.738	0.735	0.907	1.000	−0.004	−0.086	0.079	−0.105	0.097
29	If I don't have clear directions for analyzing data, I am not sure how to choose an appropriate analysis method.	0.147	0.197	0.151	1.000	0.063	−0.019	0.146	−0.064	0.190
30	Physics experiments contribute to the growth of scientific knowledge	0.955	0.943	0.286	1.000	−0.046	−0.128	0.037	−0.167	0.076

APPENDIX E: MANN WHITNEY U RESULTS FOR INDIVIDUAL ITEMS

Throughout this work, we base our findings on the AT/BT course changes of individual E-CLASS on an ANCOVA analysis using a linear regression. Since treating Likert-scale data as interval is contended, especially when looking at single items, we conducted an additional analysis based on Mann Whitney U tests presented here. We find that the Mann Whitney U results are slightly less conservative than the linear regression, but empirically equivalent to the ANCOVA analysis for the conclusions we draw in this work. However, the Mann Whitney U analysis requires more inference to interpret the results, as one cannot directly compare the pre- and post-test effect sizes.

1. Mann Whitney U analysis on individual items for women

The Mann Whitney U analysis, shown in Table XII, finds items 16, 17, and 29 significantly different in the AT and BT post-test scores and none of the items significantly different for the pretest scores. The ANCOVA analysis also found items 16, 17, and 29 to be the only significant items when controlling for the pretest scores.

2. Mann Whitney U analysis on individual items for men

The Mann Whitney U analysis, shown in Table XIII, finds items 5, 6, 14, 16, and 23 significantly different in the AT and BT post-test scores and items 4, 16, 21, and 22 significantly different for the pretest scores. The ANCOVA analysis found only items 5, 6, 14, 16, and 23 to be significant when controlling for the pretest scores. This differs from the ANCOVA analysis for items 4, 21, and 22 and potentially item 16. However, we can look at the effect sizes for each of these items to explain these differences:

Item 4. The Mann Whitney U test notes a significant difference in item 4 for the pretest. When we look at the effect sizes we see that the AT course performs higher in both the pre and post test with $g_{pre} = 0.19$ and $g_{post} = 0.13$. It is likely when we run the ANCOVA and control for

TABLE XII. E-CLASS items on the pretest and post-test compared between BT and AT scores for women using a Mann Whitney U test. The p value has been adjusted using the Holm-Bonferonni correction. A positive effect size, in both pretest and post-test, indicates that the AT course had higher mean scores than the BT course. All effect sizes in this table have Ryan-Holm step-down Bonferroni corrected $CI'_{95\%} \leq 0.21$.

		Pretest		Post-test	
#	Item	p value'	Effect size, g	p value'	Effect size, g
1	When doing an experiment, I try to understand how the experimental setup works.	1	0.001	1	−0.03
2	If I wanted to, I think I could be good at doing research.	0.2984	−0.052	1	0.04
3	When doing a physics experiment, I don't think much about sources of systematic error.	1	−0.019	1	−0.10
4	If I am communicating results from an experiment, my main goal is to have the correct sections and formatting.	0.1005	0.053	1	−0.06
5	Calculating uncertainties usually helps me understand my results better.	1	0.255	0.6498	0.18
6	Scientific journal articles are helpful for answering my own questions and designing experiments	1	0.147	1	0.08
7	I don't enjoy doing physics experiments.	0.2348	−0.017	0.1509	0.18
8	When doing an experiment, I try to understand the relevant equations.	1	−0.025	1	0.09
9	When I approach a new piece of lab equipment, I feel confident I can learn how to use it well enough for my purposes.	0.9191	0.047	1	0.10
10	Whenever I use a new measurement tool, I try to understand its performance limitations.	1	−0.078	1	0.09
11	Computers are helpful for plotting and analyzing data.	1	0.038	1	−0.03
12	I don't need to understand how the measurement tools and sensors work in order to carry out an experiment.	0.1378	0.079	1	0.03
13	If I try hard enough I can succeed at doing physics experiments.	1	−0.076	1	0.01
14	When doing an experiment I usually think up my own questions to investigate.	1	0.156	0.2564	0.17
15	Designing and building things is an important part of doing physics experiments.	1	0.039	1	0.04
16	The primary purpose of doing a physics experiment is to confirm previously known results.	0.7664	0.155	0.0002	0.31
17	When I encounter difficulties in the lab, my first step is to ask an expert, like the instructor.	1	0.057	0.0002	0.31
18	Communicating scientific results to peers is a valuable part of doing physics experiments.	1	0.059	1	0.11
19	Working in a group is an important part of doing physics experiments.	1	0.093	1	0.09
20	I enjoy building things and working with my hands.	1	−0.007	1	0.06
21	I am usually able to complete an experiment without understanding the equations and physics ideas that describe the system I am investigating.	0.7812	0.105	1	0.13
22	If I am communicating results from an experiment, my main goal is to make conclusions based on my data using scientific reasoning.	1	−0.069	1	−0.12
23	When I am doing an experiment, I try to make predictions to see if my results are reasonable.	1	0.188	0.1420	0.15
24	Nearly all students are capable of doing a physics experiment if they work at it.	1	−0.017	1	−0.01

(Table continued)

TABLE XII. *(Continued)*

#	Item	Pretest		Post-test	
		p value'	Effect size, g	p value'	Effect size, g
25	A common approach for fixing a problem with an experiment is to randomly change things until the problem goes away.	1	0.004	1	−0.10
26	It is helpful to understand the assumptions that go into making predictions.	1	0.019	1	0.02
27	When doing an experiment, I just follow the instructions without thinking about their purpose.	1	0.021	1	0.09
28	I do not expect doing an experiment to help my understanding of physics.	0.9361	−0.004	0.8498	0.00
29	If I don't have clear directions for analyzing data, I am not sure how to choose an appropriate analysis method.	1	0.063	0.0164	0.23
30	Physics experiments contribute to the growth of scientific knowledge.	1	−0.046	1	−0.06

TABLE XIII. E-CLASS items on the pretest and post-test compared between BT and AT scores for men using a Mann Whitney U test. The p value has been adjusted using the Holm-Bonferonni correction. A positive effect size, in both pretest and post-test, indicates that the AT course had higher mean scores than the BT course. All effect sizes in this table have Ryan–Holm step-down Bonferroni corrected $CI'_{95\%} \leq 0.14$.

#	Item	Pretest		Post-test	
		p value'	Effect size, g	p value'	Effect size, g
1	When doing an experiment, I try to understand how the experimental setup works.	1	0.050	1	0.010
2	If I wanted to, I think I could be good at doing research.	1	0.040	1	−0.030
3	When doing a physics experiment, I don't think much about sources of systematic error.	1	−0.020	1	−0.020
4	If I am communicating results from an experiment, my main goal is to have the correct sections and formatting.	0.0004	0.190	0.0686	0.130
5	Calculating uncertainties usually helps me understand my results better.	1	−0.010	0.0000	0.250
6	Scientific journal articles are helpful for answering my own questions and designing experiments	0.4081	0.080	0.0023	0.170
7	I don't enjoy doing physics experiments.	0.0852	0.120	1	0.030
8	When doing an experiment, I try to understand the relevant equations.	1	0.020	1	−0.020
9	When I approach a new piece of lab equipment, I feel confident I can learn how to use it well enough for my purposes.	1	−0.020	1	0.040
10	Whenever I use a new measurement tool, I try to understand its performance limitations.	1	−0.020	1	−0.080
11	Computers are helpful for plotting and analyzing data.	1	0.050	1	0.070
12	I don't need to understand how the measurement tools and sensors work in order to carry out an experiment.	1	−0.020	1	0.070
13	If I try hard enough I can succeed at doing physics experiments.	1	0.030	1	−0.070
14	When doing an experiment I usually think up my own questions to investigate.	1	0.060	0.0020	0.170

(Table continued)

TABLE XIII. *(Continued)*

#	Item	Pretest		Post-test	
		p value'	Effect size, g	p value'	Effect size, g
15	Designing and building things is an important part of doing physics experiments.	1	−0.040	0.8996	0.020
16	The primary purpose of doing a physics experiment is to confirm previously known results.	0.0008	0.180	0.0001	0.210
17	When I encounter difficulties in the lab, my first step is to ask an expert, like the instructor.	0.2608	0.110	0.5184	0.100
18	Communicating scientific results to peers is a valuable part of doing physics experiments.	1	0.050	0.5103	0.070
19	Working in a group is an important part of doing physics experiments.	1	0.060	0.2523	0.110
20	I enjoy building things and working with my hands.	1	−0.060	1	−0.040
21	I am usually able to complete an experiment without understanding the equations and physics ideas that describe the system I am investigating.	0.0026	−0.170	1	0.050
22	If I am communicating results from an experiment, my main goal is to make conclusions based on my data using scientific reasoning.	0.0106	−0.180	0.3602	−0.110
23	When I am doing an experiment, I try to make predictions to see if my results are reasonable.	1	−0.040	0.0009	0.170
24	Nearly all students are capable of doing a physics experiment if they work at it.	1	0.070	1	0.010
25	A common approach for fixing a problem with an experiment is to randomly change things until the problem goes away.	1	−0.030	1	−0.010
26	It is helpful to understand the assumptions that go into making predictions.	0.9742	−0.020	1	0.020
27	When doing an experiment, I just follow the instructions without thinking about their purpose.	1	−0.030	1	0.010
28	I do not expect doing an experiment to help my understanding of physics.	1	0.060	1	0.010
29	If I don't have clear directions for analyzing data, I am not sure how to choose an appropriate analysis method.	1	0.030	1	0.070
30	Physics experiments contribute to the growth of scientific knowledge.	1	0.000	1	−0.040

pretest scores this item is no longer significant because the change is happening in the same direction.

Item 16. The Mann Whitney U test notes a significant difference in item 16 for the pretest *and* post-test. When we look at the effect sizes we see that the AT course performs higher in both the pre and post test with $g_{\text{pre}} = 0.18$ and $g_{\text{post}} = 0.21$. The ANCOVA also found this item to have a significant change between BT and AT with a small effect size of $g = 0.16$ when calculating g based on the estimated marginal mean. This is an example of the utility of the ANCOVA analysis and controlling for the pretest scores.

Item 21. The Mann Whitney U analysis finds a significant difference in the pretest scores for item 21 where the ANCOVA analysis does not find this item to have a significant change from BT to AT. Unlike item 4 where the direction of change in pretest and post-test scores may have cancelled each other out, for item 21 the AT course performs significantly *worse* than the BT course in pretest *and* the AT course has a higher mean post-test score than the BT course. However, the pretest scores only have a small effect size of $g_{pre} = -0.17$. This is why we note that the ANCOVA analysis is slightly more conservative than the Mann Whitney U.

Item 22. The Mann Whitney U test finds significant a difference in both the pretest for in item 22. However, this time, like item 4, the direction of change in pretest and post-test scores likely cancelled each other out when controlling for the pretest scores in the ANCOVA.

[1] K. Rainey, M. Dancy, R. Mickelson, E. Stearns, and S Moller, Race and gender differences in how sense of belonging influences decisions to major in stem, Int. J. STEM Educ. **5**, 10 (2018).

[2] G. Trujillo and K. D. Tanner, Considering the role of affect in learning: Monitoring students' self-efficacy, sense of belonging, and science identity, CBE Life Sci Educ. **13**, 6 (2014).

[3] E. Brewe, L. H. Kramer, and G. E. O'Brien, Changing participation through formation of student learning communities, AIP Conf. Proc. **1289**, 85 (2010).

[4] I. Rodriguez, R. M. Goertzen, E. Brewe, and L. H. Karmer, Developing a physics expert identity in a biophysics research group, Phys. Rev. ST Phys. Educ. Res. **11**, 010116 (2015).

[5] L. Chen, S. Xu, H. Xiao, and S. Zhou, Variations in students' epistemological beliefs towards physics learning across majors, genders, and university tiers, Phys. Rev. Phys. Educ. Res. **15**, 010106 (2019).

[6] Qian Gaoyin and Donna E. Alvermann, Relationship between epistemological beliefs and conceptual change learning, Read. Writ. Q. **16**, 59 (2000).

[7] B. K. Hofer and P. R. Pintrich, The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning, Rev. Educ. Res. **67**, 88 (1997).

[8] AAPT committee on Laboratories, *AAPT Recommendations for the Undergraduate Physics Laboratory Curriculum* (AAPT, College Park, MD, 2014).

[9] D. Doucette, R. Clark, and C. Singh, Hermione and the secretary: How gendered task division in introductory physics labs can disrupt equitable learning, Eur. J. Phys. **41**, 035702 (2020).

[10] Benjamin M. Zwickl, Takako Hirokawa, Noah 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).

[11] Bethany 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).

[12] Brian A. Danielak, Ayush Gupta, and Andrew Elby, Marginalized identities of sense-makers: Reframing engineering student retention, J. Engin. Educ. **103**, 8 (2014).

[13] Michael Enman and Judy Lupart, Talented female students' resistance to science: An exploratory study of post-secondary achievement motivation, persistence, and epistemological characteristics, High Ability Studies **11**, 161 (2000).

[14] Katherine N. Quinn, Michelle M. Kelley, Kathryn L. McGill, Emily M. Smith, Zachary Whipps, and N. G. Holmes, Group roles in unstructured labs show inequitable gender divide, Phys. Rev. Phys. Educ. Res. **16**, 010129 (2020).

[15] H. Lewandowski, D. Bolton, and B. Pollard, Initial impacts of the transformation of a large introductory lab course focused on developing experimental skills and expert epistemology, *PER Conf. 2018, Washington, D.C.*, 10.1119/perc.2018.pr.Lewandowski

[16] R. Gutiérrez and E. Dixon-Román, *An Introduction to Statistical Concepts* (Springer, New York, 2011).

[17] R. Gutierrez, A "gap-gazing" fetish in mathematics education? Problematizing research on the achievement gap, J. Res. Math. Educ. **39**, 357 (2008).

[18] Elaine Seymour, Anne-Barrie Hunter, R. P. Harper, and D. G. Holland, Talking About Leaving Revisited: Persistence, Relocation, and Loss in Undergraduate STEM Education, edited by Elaine Seymour and Anne-Barrie Hunter (Springer, Cham, Switzerland, 2019), 10.1007/978-3-030-25304-2.

[19] O. Lee, Equity implications based on the conceptions of science achievement in major reform documents, Rev. Educ. Res. **69**, 83 (1999).

[20] Anna T. Danielsson, Exploring woman university physics students "doing gender" and "doing physics", Gender Educ. **24**, 25 (2012).

[21] Zahra Hazari, Eric Brewe, Renee Michelle Goertzen, and Theodore Hodapp, The importance of high school physics teachers for female students' physics identity and persistence, Phys. Teach. **55**, 96 (2017).

[22] Adrienne L. Traxler, Ximena C. Cid, Jennifer Blue, and Ramón Barthelemy, Enriching gender in physics education research: A binary past and a complex future, Phys. Rev. Phys. Educ. Res. **12**, 020114 (2016).

[23] Adrian Madsen, McKagan, B. Sarah, and Eleanor C. Sayre, Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap?, Phys. Rev. ST Phys. Educ. Res. **9**, 020121 (2013).

[24] M. Lorenzo, Catherine Hirshfeld Crouch, and E. Mazur, Reducing the gender gap in the physics classroom, Am. J. Phys. **74**, 118 (2006).

[25] J. Docktor and K. Heller, Gender differences in both force concept inventory and introductory physics performance, AIP Conf. Proc. **1064**, 15 (2008).

[26] 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).

[27] I. Rodriguez, G. Potvin, and L. H. Kramer, How gender and reformed introductory physics impacts student success in advanced physics courses and continuation in the physics major, Phys. Rev. Phys. Educ. Res. **12**, 020118 (2016).

[28] J. Hyde and M. Linn, Gender similarities in mathematics and science, Science **314**, 599 (2006).

[29] C. M. Steele, A threat in the air: How stereotypes shape intellectual identity and performance, Am. Psychol. **52**, 613 (1997).

[30] Alexandru Maries, Nafis I. Karim, and Chandralekha Singh, Is agreeing with a gender stereotype correlated with the performance of female students in introductory physics?, Phys. Rev. Phys. Educ. Res. **14**, 020119 (2018).

[31] S. Cheryan, J. O. Siy, M. Vichayapai, B. J. Drury, and S. Kim, Do female and male role models who embody stem stereotypes hinder women's anticipated success in stem?, Social Psychol. Personality Sci. **2**, 656 (2011).

[32] H. Mendick, Mathematical stories: Why do more boys than girls choose to study mathematics at as-level in England?, Br. J. Sociol. Educ. **26**, 235 (2005).

[33] T. Mujtaba and M. J. Reiss, Inequality in experiences of physics education: Secondary school girls' and boys' perceptions of their physics education and intentions to continue with physics after the age of 16, Int. J. Sci. Educ. **35**, 1824 (2012).
[34] H. B. Carlone, Innovative science within and against a culture of "achievement", Sci. Educ. **87**, 307 (2003).
[35] Valerie Otero, Steven Pollock, and Noah Finkelstein, A physics department's role in preparing physics teachers: The Colorado Learning Assistant Model, Am. J. Phys. **78**, 1218 (2010).
[36] Jacob T. Stanley and H. J. Lewandowski, Lab notebooks as scientific communication: Investigating development from undergraduate courses to graduate research, Phys. Rev. Phys. Educ. Res. **12**, 020129 (2016).
[37] Jacob T. Stanley and H. J. Lewandowski, Recommendations for the use of notebooks in upper-division physics lab courses, Am. J. Phys. **86**, 45 (2018).
[38] Dimitri R. Dounas-Frazer and H. J. Lewandowski, The modelling framework for experimental physics: Description, development, and applications, Eur. J. Phys. **39**, 064005 (2018).
[39] B. M. Zwickl, D. Hu, N. Finkelstein, and H. J. Lewandowski, Model-based reasoning in the physics laboratory: Framework and initial results, Phys. Rev. Phys. Educ. Res. **11**, 020113 (2015).
[40] H. J. Lewandowski, Benjamin Pollard, and Colin G. West, Using custom interactive video prelab activities in a large introductory lab course, *PER Conf. 2019, Provo, UT*, 10.1119/perc.2019.pr.Lewandowski.
[41] Bethany R. Wilcox and H. J. Lewandowski, Students' views about the nature of experimental physics. Phys. Rev. Phys. Educ. Res. **13**, 020110 (2017).
[42] John M. Aiken and H. J. Lewandowski, Data sharing model for physics education research using the 70000 response Colorado Learning Attitudes about Science Survey for experimental physics dataset, Phys. Rev. Phys. Educ. Res. **17**, 020144 (2021).
[43] B. R. Wilcox and H. J. Lewandowski, A summary of research-based assessment of students' beliefs about the nature of experimental physics, Am. J. Phys. **86**, 212 (2018).
[44] Bethany R. Wilcox, Benjamin M. Zwickl, Robert D. Hobbs, John M. Aiken, Nathan M. Welch, and H. J. Lewandowski, Alternative model for administration and analysis of research-based assessments, Phys. Rev. Phys. Educ. Res. **12**, 010139 (2016).
[45] Nidhal Sulaiman, Benjamin Pollard, and H. J. Lewandowski, Impact on students' views of experimental physics from a large introductory physics lab course, *PER Conf. 2020, virtual conference*, 10.1119/perc.2020.pr.Sulaiman.
[46] J. Keller, Blatant stereotype threat and women's math performance: Self-handicapping as a strategic means to cope with obtrusive negative performance expectations, Sex Roles, **47**, 193 (2002).
[47] A. Miyake, L. E. Kost-Smith, N. D. Finkelstein, S. J. Pollock, L. Cohen, G, and T. A. Ito, Reducing the gender achievement gap in college science: A classroom study of values affirmation, Science **330**, 1234 (2010).
[48] D. L. Hahs-Vaughn and R. G. Lomax, *An Introduction to Statistical Concepts* (Routledge, New York, 2020).
[49] R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2020).
[50] J. Cohen, The statistical power of abnormal-social psychological research: A review, J. Abnormal Social Psychol. **65**, 145 (1962).
[51] L. V. Hedges, Distribution theory for Glass's estimator of effect size and related estimators, J. Educ. Stat. **6**, 107 (1981).
[52] B. Thompson, What future quantitative social science research could look like confidence intervals for effect sizes, Educ. Res. **31**, 25 (2002).
[53] B. Thompson, Effect sizes, confidence intervals, and confidence intervals for effect sizes, Psychol. Schools **44**, 423 (2008).
[54] J. Ludbrook, Multiple inferences using confidence intervals, Clinical Experimental Pharmacol. Physiol. **27**, 212 (2000).
[55] C. O. Fritz, P. E. Morris, and Jennifer J. Richler, Effect size estimates: Current use, calculations, and interpretation, J. Exper. Psychol. General **141**, 2 (2012).
[56] J. Carifio and R. Perla, Resolving the 50-year debate around using and misusing Likert scales, Med. Educ. **42**, 1150 (2008).
[57] Patrick E. McKnight and Julius Najab, Mann-Whitney U test, *The Corsini Encyclopedia of Psychology* (John Wiley & Sons, Ltd, New York, 2010), p. 1, 10.1002/9780470479216.corpsy0524.
[58] Russell Lenth, Henrik Singmann, Jonathon Love, Paul Buerkner, and Maxime Herve, Emmeans: Estimated marginal means, aka least-squares means, R package version **1**, 3 (2018), https://github.com/rvlenth/emmeans.
[59] Lauren E. Kost, Steven J. Pollock, and Noah D. Finkelstein, Characterizing the gender gap in introductory physics, Phys. Rev. ST Phys. Educ. Res. **5**, 010101 (2009).
[60] B. Pollard and H. J. Lewandowski, Transforming a large introductory lab course: Impacts on views about experimental physics, *PER Conf. 2018, Washington, DC*, 10.1119/perc.2018.pr.Pollard.
[61] Emily M. Smith, Martin M. Stein, and N. G. Holmes, How expectations of confirmation influence students' experimentation decisions in introductory labs, Phys. Rev. Phys. Educ. Res. **16**, 010113 (2020).
[62] Dehui Hu, Benjamin M. Zwickl, Bethany R. Wilcox, and H. J. Lewandowski, Qualitative investigation of students' views about experimental physics, Phys. Rev. Phys. Educ. Res. **13**, 020134 (2017).
[63] Alec Sithole, Edward T. Chiyaka, Peter McCarthy, Davison M. Mupinga, Brian K. Bucklein, and Joachim Kibirige, Student attraction, persistence and retention in STEM programs: Successes and continuing challenges, Higher Educ. Studies **7**, 46 (2017).
[64] Trent W. Maurer, Deborah Allen, Delena Bell Gatch, Padmini Shankar, and Diana Sturges, A comparison of student academic motivations across three course disciplines, J. Scholarship Teaching Learning **13**, 77 (2013), https://scholarworks.iu.edu/journals/index.php/josotl/article/view/2153.

[65] A. T. Danielsson and C. Linder, Learning in physics by doing laboratory work: Towards a new conceptual framework, Gender Educ. **21**, 129 (2009).

[66] Z. Yasemin Kalender, Emily Stump, Katelynn Hubenig, and N. G. Holmes, Restructuring physics labs to cultivate sense of student agency, Phys. Rev. Phys. Educ. Res. **17**, 020128 (2021).

[67] Dimitri R. Dounas-Frazer, Jacob T. Stanley, and H. J. Lewandowski, Student ownership of projects in an upper-division optics laboratory course: A multiple case study of successful experiences., Phys. Rev. Phys. Educ. Res. **13**, 020136 (2017).

[68] Erika M. Nadile, Keonti D. Williams, Nicholas J. Wiesenthal, Katherine N. Stahlhut, Krystian A. Sinda, Christopher F. Sellas, Flor Salcedo, Yasiel I. Rivera Camacho, Shannon G. Perez, Meagan L. King, Airyn E. Hutt, Alyssa Heiden, George Gooding, Jomaries O. Gomez-Rosado, Sariah A. Ford, Isabella Ferreira, Megan R. Chin, William D. Bevan-Thomas, Briana M. Barreiros, Emilie Alfonso, Yi Zheng, and Katelyn M. Cooper, Gender differences in student comfort voluntarily asking and answering questions in large-enrollment college science courses, J. Microbiol. Biol. Educ. **22**, e00100 (2021).