Editors' Suggestion

Skills-focused lab instruction improves critical thinking skills and experimentation views for all students

Cole Walsh,[1] H. J. Lewandowski,[2,3] and N. G. Holmes[1,*]

[1]*Laboratory of Atomic and Solid State Physics, Cornell University, Ithaca, New York 14853, USA*
[2]*Department of Physics, University of Colorado Boulder, Boulder, Colorado 80309, USA*
[3]*JILA, National Institute of Standards and Technology and the University of Colorado, Boulder, Colorado 80309, USA*

(Received 12 January 2022; accepted 14 March 2022; published 11 April 2022)

Instructional labs are fundamental to an undergraduate physics curriculum, but their possible learning goals are vast with limited evidence to support any particular goal. In this study, we evaluate the efficacy of labs with different goals and structures on students' critical thinking skills and views about experimentation, using an extensive database of survey responses from over 20 000 students at over 100 institutions. Here, we show that labs focused on developing experimentation skills improve students' critical thinking skills and experimentation views compared to labs focused on reinforcing lecture concepts. We further demonstrate the positive impacts of skills-based labs over concepts-based labs on these outcomes across students' gender and race or ethnicity. Our analysis also shows that activities to support students' decision making and communication explain over one-half and one-third of the effect of skills-based labs on students' critical thinking skills and experimentation views, respectively, while modeling activities have only a small effect on performance.

DOI: 10.1103/PhysRevPhysEducRes.18.010128

I. INTRODUCTION

Instructional labs make up an important part of the undergraduate physics curriculum, with the opportunity to engage students in the practices of experimental physics and develop their technical, communication, and critical thinking skills [1]. There are myriad ways that labs can be structured, consistent with a wide range of learning goals [2–6]. However, literature on lab instruction demonstrates a lack of consensus on the desired goals of lab instruction [2,7], with many labs focusing on demonstrating or reinforcing canonical theories or phenomena, rather than seeking to develop students' skills [8]. The lack of consensus on goals for labs has been attributed to a lack of evidence about the efficacy of labs for either goal [2,9,10]. More and more research has begun evaluating the efficacy of different lab programs on students' critical thinking skills and views, but a comprehensive study across multiple curricula is needed.

Research has suggested that labs with an explicit goal of reinforcing concepts also taught in lecture have little to no impact on students' conceptual understanding [11–16]. These types of labs (reinforcing concepts from lecture) have also been shown to deteriorate student's attitudes and beliefs about experimental physics [14,17,18]. In contrast, labs designed to teach experimentation skills have been found to improve students' experimentation skills [14,15,19–21] and their attitudes and beliefs about experimental physics [14,17,18], with no measurable impact to their conceptual understanding [14–16].

In this paper, we probe an open area of research that looks to evaluate the efficacy of labs with different goals. We examine how labs from over 100 institutions impact students' critical thinking skills and experimentation views, and, more importantly, evaluate whether the impact is consistent for students of different genders and races or ethnicities. Additionally, we evaluate what types of instructional activities may explain the impacts on students' outcomes. Ultimately, the analysis demonstrates the universally positive impact of skills-based labs compared with concepts-based labs on these assessments for all demographics of students, due in part by their increased use of activities that target student decision making and communication.

A. How does lab type impact subpopulations of students?

Several previous studies have looked at the effects of labs with different goals on overall student performance. For example, research has generally found benefits of labs

*ngholmes@cornell.edu

aiming to develop students' experimentation skills, looking at whole cohorts of students. Many studies in physics education research, however, have found (or suggested) that instructional outcomes may differ for different subpopulations of students [22–24]. In labs, specifically, researchers have found that men held more positive attitudes and beliefs about experimental physics than women, on average, and that the size of this difference was largely maintained following instruction [25]. Similar differences have been found on measures of student performance on lab-specific outcomes, such as one study that found men outperformed women on a data handling diagnostic, on average, and that the difference widened following instruction [26]. Analysis of students' grades, however, found no consistent difference between men's and women's grades in lab courses, despite consistent differences in lecture courses [27].

Some of these differences may be a result of how different students experience or are able to participate in the instruction. For example, previous work found that men and women participate in lab activities and roles differently [26,28–33]. This division of tasks, however, may vary depending on the goal or type of lab instruction [30]. In a more open-ended lab that aimed to develop students' experimentation skills, researchers found that there were more roles available to students and that the division of roles along gender lines was more distinctive than in a traditional lab aiming to reinforce concepts [30].

Alternatively, the differences (or lack thereof) may be inherent to the pedagogical structures of laboratory courses in general. For example, the assessment and grading structures in lab courses differ significantly from those of lecture courses [27], which likely impacts the ways students approach the courses [34]. The prolific use of group work [8], often such that students are assessed at the group level, may also be a factor. Altogether, this literature raises the question: How do labs with different purposes impact the attitudes and experimentation skills of students from different subpopulations?

B. What pedagogical features of different lab types lead to outcomes?

Given the apparent benefits of labs aiming to develop students' experimental physics skills, the next question is: why? Given that it is the implementation, rather than the goal, of the lab that impacts students' learning, through what mechanisms do labs with different goals impact students' performance?

To answer this question, we explored the types of skills the instructors indicated focusing on through their lab activities. For example, lab courses may focus on data analysis and uncertainty, experimental design, modeling, or communication skills, among others [3]. Lab courses in this dataset almost ubiquitously included activities associated with data analysis and uncertainty [8], restricting us from drawing any conclusions about the impacts of these skills on student outcomes. Three other types of skills, however, were more variable in the dataset.

First, many courses focused on developing students' skills around making decisions about an experiment, such as to choose a research question or design aspects of the experimental procedures. Pedagogically, supporting students' decision making requires opportunities and support for student agency [35]. Labs that support student agency and choice have been shown to improve students' engagement [36], attitudes and beliefs [14,37,38], and engagement in experimental physics practices [15,19,35].

Second, a fundamental aspect of experimental physics is learning to generate and test knowledge about physics through modeling activities. A focus on modeling has been seen in high school physics [39], introductory physics instruction [40], and upper-division physics labs [41]. A modeling focus has been shown to engage students in behaviors that align with more expertlike thinking and reasoning in the lab [42,43] and support students' attitudes towards physics [44].

Finally, communication skills are a significant focus of many lab courses [45], as students learn to use lab notebooks [46], write lab reports or mock journal articles [47,48], or present their work [48]. Communication skills may also relate to teamwork skills [48], which may in turn develop classroom community and foster students' sense of belonging [49].

C. Research questions

To measure students' experimentation views and critical thinking skills, we analyzed data from two previously validated assessments: the Colorado Learning Attitudes about Science Survey for Experimental Physics (E-CLASS) [50] and the Physics Lab Inventory of Critical thinking (PLIC) [21]. Together, these two instruments have been used in hundreds of courses in over 100 different institutions, totaling in responses from over 20 000 students around the world. Our research questions in this study are threefold:

RQ1 How do labs with different purposes affect the E-CLASS and PLIC scores for different subpopulations of students?

RQ2 What pedagogical features are characteristic of labs with different intended purposes?

RQ3 How do pedagogical features of labs affect students' E-CLASS and PLIC scores?

We distinguish three different types of labs based on their overall purpose: (a) labs that aim to reinforce material from lecture (concepts-based labs), (b) labs that aim to develop students' experimentation skills (skills-based labs), and (c) labs that aim to do both (mixed labs).

To answer RQ1, we evaluate students' outcomes through a lens of equity of individuality [24,51–53]: "Equity of individuality is achieved when an intervention improves the outcomes of students from marginalized

groups" [[24], p. 40]. With this definition of equity, we do not explicitly examine whether achievement gaps exist between groups of students, nor whether these potential gaps are closed, maintained, or widened following instruction. We instead examine whether labs with different purposes provide positive outcomes for all groups of students.

For RQ2 and RQ3, we consider pedagogical features that relate to decision-making, modeling, and communication activities that students engage in during labs. For RQ2, we examine how instructors' intended goals for their labs align with the prevalence of these activities in their labs, an indication of their pedagogical choices. For RQ3, we evaluate the effects of these pedagogical features on students' E-CLASS and PLIC scores to ascertain the role of these features in developing students' experimentation views and critical thinking skills in physics labs. By addressing these two research questions, we begin to explore how instructors' intended goals for their labs manifest in their pedagogical choices and the impacts of these pedagogical choices.

II. METHODS

In this section, we provide an overview of the data and analysis methods used in this study. Additional details can be found in Appendix A.

A. Data sources

We measured students' critical thinking skills and views about experimental physics through two instruments.

The E-CLASS aims to measure students' personal views about experimental physics in their lab class by evaluating the degree to which students agree or disagree with statements about experimental physics [54]. The instrument consists of 30 five-point Likert items and students are scored based on how well their responses align with responses from expert physicists on a collapsed three-point scale: students receive one point on an item if their answer aligns with the majority of experts (e.g., the student selects "agree" or "strongly agree" when most experts selected "strongly agree" or "agree") and −1 points if their response is opposite to that selected by the majority of experts (e.g., the student selects "agree" or "strongly agree" when most experts selected "strongly disagree" or "disagree"). Neutral responses receive zero points. This scoring scheme provides a range of possible scores on the E-CLASS from −30 to 30 [25,54].

The PLIC aims to measure students' critical thinking skills in the context of experimental physics, defined here as the decision making involved in interpreting data, drawing accurate conclusions from data, comparing and evaluating models and data, evaluating methods, and deciding how to proceed in an investigation [21,55]. The PLIC consists of 10 multiple-response items and, as with the E-CLASS, students are scored according to how well their responses align with those from expert physicists. Scores on each item can range from zero to one with partial credit awarded for selecting response choices that were picked by at least 10% of experts. Possible scores on the PLIC, then, range from zero to ten.

When administering the E-CLASS or PLIC, instructors provided details about their class through a course information survey (CIS) [17,56]. The CIS asks, for example, about the course level, the number of hours students spent in lab each week, the number of instructional staff, and the main purpose of the lab (either to reinforce physics concepts, develop lab skills, or both about equally). We acknowledge that one can not isolate skills or concepts in labs entirely; models and practices are inextricably linked in experimental physics. These broad categorizations of the lab purpose, however, informs whether, primarily, the practice is in service of theory or the theory is in service of practice [11]. Because different instructors may view these characterizations differently, we also explore more tangible pedagogical variables that may more specifically characterize these categories of lab purposes. The survey also asks how often students engage in various activities in the lab, such as designing procedures, building apparatus, or working in groups. We used this last set of items to measure the amount of decision making, modeling activities, and communication activities in the labs, discussed in further detail in Appendix B.

We used data collected from only first-year courses, as labs at the beyond-first-year level tended to be more homogeneous and aligned with developing students' lab-related skills; only four beyond-first-year courses in our E-CLASS and PLIC datasets were labeled as concepts-based by instructors. We used responses to the E-CLASS from 16 409 students enrolled in 230 classes and 56 institutions, and responses to the PLIC from 4988 students enrolled in 77 classes and 28 institutions. We define a class here as a combination of course (e.g., Physics 101) and semester (e.g., Fall 2019), so a single course may administer the E-CLASS or PLIC in multiple semesters and count as multiple classes. In Table I, we provide the breakdown of institutions and classes in our datasets across institution type and the main purpose of the lab associated with the class.

TABLE I. Breakdown of the institutions and classes included in the E-CLASS and PLIC datasets.

		E-CLASS	PLIC
Institutions	2-year college	4	0
	4-year college	21	13
	Master's granting	5	1
	Ph.D. granting	26	14
Lab type	Concepts based	51	18
	Skills based	56	32
	Mixed	123	27

TABLE II. Demographic breakdown of students in the E-CLASS and PLIC datasets across lab type. Racial or ethnic groups were not considered mutually exclusive and so numbers may not sum to the total number of students in the dataset.

	E-CLASS				PLIC			
Student-level variables	Full sample	Concepts based	Skills based	Mixed	Full sample	Concepts based	Skills based	Mixed
All	16 409	3209	3823	9377	4988	1838	2229	921
Gender								
Man	9236	1604	2234	5398	2762	1065	1168	529
Nonbinary	172	29	48	95	51	11	29	11
Woman	6626	1489	1482	3655	2140	753	1013	374
Unknown	375	87	59	229	35	9	19	7
Race or ethnicity								
American Indian	152	17	22	113	63	22	29	12
Asian	4060	603	717	2740	1609	664	731	214
Black	1072	538	100	434	232	51	134	47
Hispanic	1508	279	383	846	457	141	235	81
Native Hawaiian	153	25	30	98	25	10	8	7
White	9394	1752	2487	5155	2899	1063	1234	602
Other	406	71	86	249	202	51	102	49
Unknown	1061	193	264	604	137	44	68	25

Both the E-CLASS and PLIC are administered to students prior to lab instruction (pretest) and following the conclusion of lab instruction for the semester (post-test). We collected students' self-reported gender, race or ethnicity, and academic major information at the end of the surveys. Table II gives a breakdown of the students in our datasets by students' self-identified gender and race or ethnicity and by the type of lab in which the student was enrolled. For both instruments, students had the option of not disclosing any demographic information, in which case we categorized their gender or race or ethnicity as *unknown*. We kept these students in our dataset to maintain statistical power and recognize all students who completed the assessments whether or not they were comfortable disclosing demographic information. We also chose not to collapse demographic characteristics (such as by grouping students into majority or underrepresented minority categories) to more accurately evaluate the possible effects of lab instruction on different subgroups of students [57].

B. RQ1: Effects of different lab types on scores for different subpopulations of students

To address RQ1, we performed mixed-model regression analysis to evaluate the impact of lab type on student outcomes for different subpopulations of students. We used two-level linear mixed models with institutions as random effects to account for students being nested within institutions in our datasets. The institutional random effects help to account for potential systematic differences between institutions, such as prior preparation or instruction in nonlab components of the courses (for a review of linear mixed models for physics education research datasets, see Ref. [58]). Unconditional models—models with no fixed effects, but with institution as a random effect—indicated that 5.1% of the variation in E-CLASS post-test scores and 7.1% of the variation in PLIC post-test scores could be explained by institution-level differences alone.

Separately for the E-CLASS and PLIC, we fit models with *post-test score* as the dependent variable and main effects for *lab type*, *pretest score*, *gender*, *race or ethnicity*, and *major*. We additionally included interaction terms between *lab type* and *gender*, and *lab type* and *race or ethnicity*. This model allowed us to investigate whether the effect of lab type on students' post-test scores differed across student demographic groups. We used a lens of equity of individuality [24,51], whereby we evaluated the post-test scores in each intervention (i.e., lab type) separately for each demographic group. Using pretest score as a main effect variable serves to account for differences in students' incoming preparation.

C. RQ2: Pedagogical features that characterize different types of labs

We used a confirmatory factor analysis (CFA) to construct a measurement model for the amount of decision making, modeling, and communication activities in labs using a combined dataset of instructors' responses to both the E-CLASS and PLIC CIS. To avoid overweighting courses where instructors administered both the E-CLASS and PLIC in the same course or either assessment in multiple classes across semesters, we included only one entry for each unique course in our dataset. We made an exception if an instructor provided different responses to the CIS for different classes, indicating that pedagogical features of their lab had changed. The dataset used in this analysis included 157 unique courses (or classes where the instructor provided different responses to the CIS in different semesters). The full measurement model used and its development are presented in Appendix B.

To address RQ2, we used Thurstone's regression method [59] and our measurement model to compute factor scores for the amount of decision making, modeling, and communication activities in each of the 157 unique courses in our dataset. We standardized factor scores to have mean zero and standard deviation 1.

D. RQ3: Effects of pedagogical variables on students' scores

To address RQ3, we further computed factor scores for all classes in both the E-CLASS and PLIC datasets using the same procedure as above. Eight classes (corresponding to two unique courses who used the PLIC multiple times) with 158 students total were missing information on the PLIC CIS such that we could not compute their factor scores, and so were removed from the dataset. We used the complete factor scores data in two-level linear mixed models similar to those described in Sec. II B (separately for the E-CLASS and for the PLIC) with post-test score as the dependent variable and controlling for the main effects of *pretest scores*, *gender*, *race or ethnicity*, and *major*. Unlike the models described in Sec. II B, we did not include *lab type* as a main effect and we did not include interaction effects between any variables.

Because of a large correlation between the factor scores for *decision making* and *communication* activities ($r = 0.93$), we fit two separate models for each of the E-CLASS and PLIC datasets. In one model, we included *decision making* and *modeling* factor scores, while in the other model we included *modeling* and *communication* factor scores. These models allowed us to determine the effect of increased decision making and communication on students' scores controlling for the amount of modeling in labs, and vice versa. We could not, however, disentangle the effects of decision making and communication on students' scores, given the significant correlation.

III. RESULTS

A. RQ1: Effects of different lab types on scores for different subpopulations of students

Full results from the fitted linear mixed models for the E-CLASS and PLIC are presented in Appendix C. In this section, we summarize the results using plots of expected post-test scores from marginal effects. Marginal effects represent the expected outcomes from our fitted models and indicate how the outcome measure (i.e., E-CLASS and PLIC post-test scores) changes with particular independent variables (i.e., lab type, gender, and race or ethnicity). In these marginal effects plots, students' pretest scores are held fixed at the mean value for each independent variable being investigated. All other variables, other than the ones plotted, are held fixed at their proportions. For example, 13.8% of students in our E-CLASS dataset intended to major in math or computer science. When calculating marginal effects from this model, we set the Math and CS variable, which is generally a binary variable (1 if the student is a math or CS major and 0 otherwise), equal to 0.138. These marginal effects should, then, be interpreted as expected post-test scores averaged across all other variables.

Figure 1 shows the expected post-test scores for the E-CLASS and the PLIC. The first panel in each row shows the effect of lab type on aggregate. Overall, on both assessments, students in skills-based labs score an average of 0.2 standard deviations higher at post-test than students in concepts-based labs, controlling for students' pretest scores, and self-reported major, gender, and race or ethnicity. Mixed labs sit between the two extremes.

The second and third panels of Fig. 1 show that, when there is sufficient precision to distinguish the groups, the pattern is consistent across student gender and race or ethnicity. That is, students from all subpopulations score higher on the instruments when participating in skills-based labs compared to concepts-based labs, again with mixed labs in the middle. The size of the effect differs for different subpopulations, however, and the sample sizes in some subpopulations are too small to distinguish the scores between lab type.

B. RQ2: Pedagogical features that characterize different types of labs

We found differences in the average factor scores across all three pedagogical variables between skills-based and concepts-based labs. For agency factor scores, this difference was about 1.44 ± 0.17 standard deviations; for modeling factor scores, this difference was about 0.51 ± 0.23 standard deviations; and for communication factor scores, this difference was about 1.58 ± 0.16 standard deviations. Smoothed density plots of the fraction of labs of each type with varying amounts of agency, modeling activities, and communication activities are shown in Fig. 2. We find that skills-based labs typically engage students in more decision making and communication activities than concepts-based labs, while mixed labs typically fall somewhere between these extremes. The amount of modeling activities in all three types of labs follow similar distributions, with mixed and skills-based labs supporting slightly more modeling than concepts-based labs, on average.

We also found that these pedagogical variables were not independent. The presence of modeling activities in labs was moderately correlated with an increase in decision making ($r = 0.41$) and communication activities ($r = 0.23$), while increased student decision making was highly correlated with increased communication activities ($r = 0.93$). These results imply that pedagogical choices made to support student decision-making in physics labs are closely associated with choices to support opportunities for student communication.

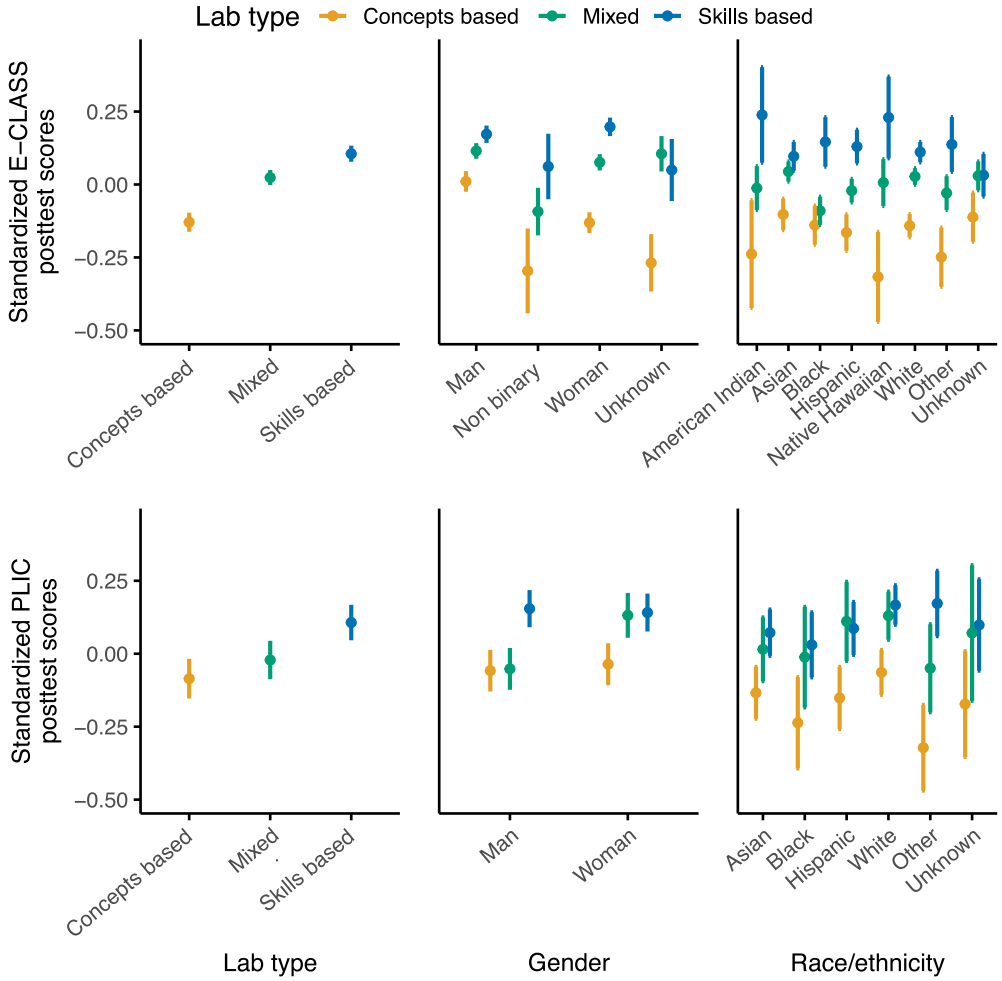

FIG. 1. Expected post-test scores (i.e., marginal effects) across lab type and student demographics. Pretest score is fixed at the mean, while all other variables are fixed at their proportions. Students score higher at post-test when enrolled in skills-based labs as compared to concepts-based or mixed labs and the pattern is consistent across student demographics (when precision can distinguish the groups). Error bars represent one standard error (68% confidence interval). We have not shown marginal effects on the PLIC for students who identified as nonbinary, an unknown gender, American Indian or Alaska Native, or Native Hawaiian or Pacific Islander due to error bars that exceed the range of the plots. These students were included in our models, however, and full results can be found in Appendix C.

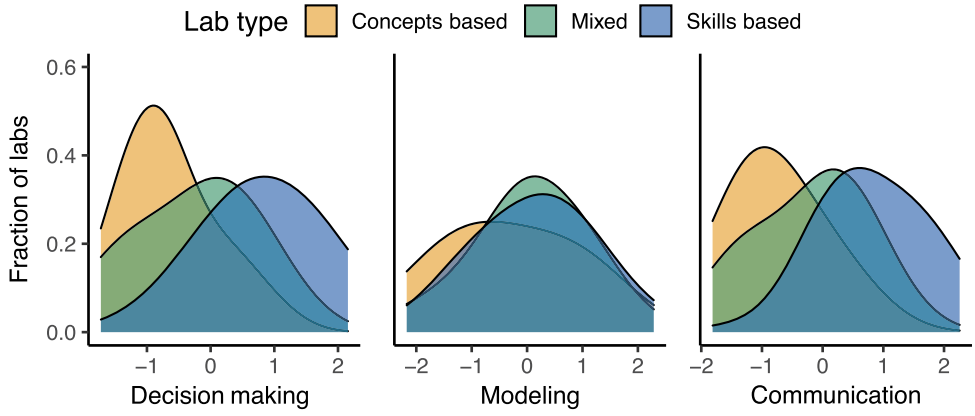

FIG. 2. Density plots of the amount of agency, modeling, and communication in labs with different intended purposes.

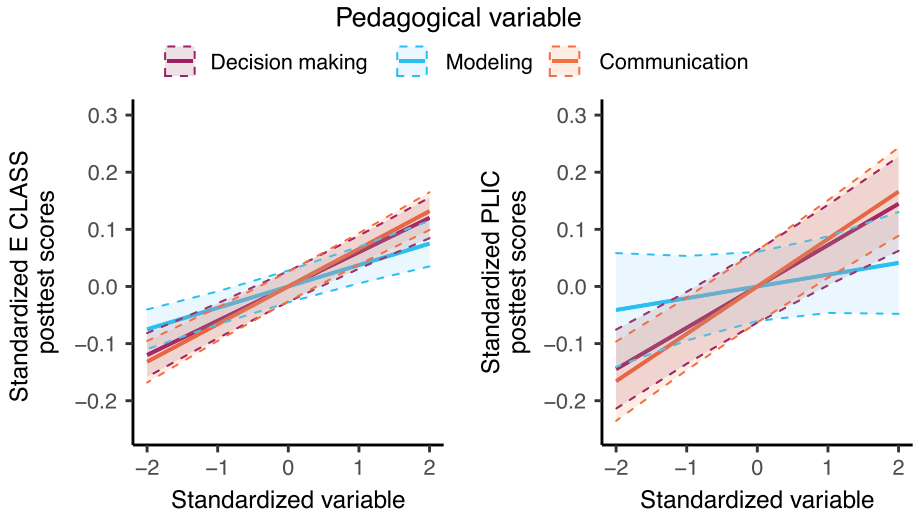

FIG. 3. Expected post-test scores (i.e., marginal effects) on the E-CLASS and PLIC as a function of time spent on decision-making, modeling, and communication activities. Post-test scores are higher in labs with more activities related to decision making and communication, with a small effect from modeling activities. The shading represents the 68% confidence intervals on the estimates.

C. RQ3: Effects of pedagogical variables on students' scores

Figure 3 shows the expected post-test scores as a function of the amount of decision-making, modeling, and communication activities in the labs for both the E-CLASS and the PLIC (Appendix C for the full results of the fitted linear mixed models). Post-test scores are higher in labs with more activities related to decision making and communication, with a small effect from modeling activities.

We also estimated the fraction of the observed effect of skills-based labs (compared to concepts-based labs) that can be explained by these pedagogical variables (either decision making, modeling, or communication activities). Full results and calculations can be found in Appendix C. We find that decision-making and communication opportunities in labs accounted for 34%–41% of the observed effect of skills-based labs on students' scores on the E-CLASS and 58%–76% of the effect on PLIC scores (corresponding to standardized effect sizes of 0.06–0.09). In contrast, the presence of modeling activities in labs accounted for less than 7% of the difference in scores between skills-based and concepts-based labs on both the E-CLASS and PLIC (standardized effect sizes of 0.03 ± 0.01 and 0.02 ± 0.03 for the E-CLASS and PLIC, respectively).

IV. DISCUSSION

Our analysis validates previous work [17] demonstrating the overall effectiveness of skills-based labs in developing students' views about experimental physics, now with a much broader dataset. We also demonstrate similar effects on students' critical thinking skills in the context of experimental physics. By simultaneously comparing student scores by lab type and demographic variables, we illustrate that this effect is consistently positive regardless of students' gender and race or ethnicity.

Although all students benefited from skills-based labs, there was still a differential impact on subpopulations of students. The difference in scores on the E-CLASS between students in skills-based versus concepts-based labs was much larger for women than men, consistent with previous work [17]. One might infer that concepts-based labs are particularly detrimental to women's views or that skills-based labs are particularly beneficial to women's views. Future work, such as through interviews and video analysis of students in labs, should further evaluate how women experience skills-based labs compared with concepts-based labs, particularly given the effects on their participation seen previously [26,28–33], and how this relates to their views of experimental physics more broadly.

This differential impact on subgroups of students in different types of labs, however, did not exist with the PLIC. The difference between students' PLIC scores in skills-based and concepts-based labs were the same regardless of students' gender. This result suggests that skills-based and concepts-based labs affect men's and women's critical thinking skills similarly, despite differences in how it affects their views. The impact of mixed labs, however, on students' PLIC scores differed for men and women in ways that cannot be explained by the dataset. Future work should evaluate possible explanations for this differential impact.

The size of the gaps between lab types for students of different reported races or ethnicities were more variable on both assessments, as was the precision with which we could measure them. Future work should continue to evaluate the impacts of different types of labs on students with different racial or ethnic identities.

We also find evidence that skills-based labs typically incorporated more activities to support decision making

and communication skills than concepts-based or mixed labs, which correlated with students' PLIC and E-CLASS performance. Modeling activities, on the other hand, had a smaller difference in prevalence between lab types and a smaller effect on students' post-test scores. The results indicate that the measures of decision making and communication activities explain most, but not all, of the variability between scores based on lab type.

The additional variability between lab type and student scores may come from limitations in our ability to measure the pedagogical activities in the courses. First, our analyses initially relied on instructors' classifications of their courses as aiming to reinforce conceptual understanding, develop lab skills, or both about equally. Individual instructors may have used different criteria for characterizing their course along these lines, as evidenced by the variability with which the courses incorporate the three pedagogical features studied. In addition, the analysis captured only the instructors' perceptions of their instruction, not necessarily what actually took place, and captured only a finite (and not exhaustive) set of activities related to these pedagogical variables. For example, the analysis does not include the role of grading, instructor feedback during class time, or the role of interactions between students, all of which may impact the enactment of these pedagogies. Analysis of instructors' course materials (e.g., syllabi, explicit learning goals, lab instructions) and the enactment of those materials in classrooms could more accurately (albeit less efficiently) determine the actual instruction carried out in the labs.

Alternatively, the additional variability may come from additional types of activities included in the instruction beyond the three variables explored here. For example, the role of grading in labs may be a critical variable. Prior work found no differences in men's and women's physics lab grades, despite consistent differences in their lecture course grades [27]. The authors attributed this result to differences in the testing and grading structures in labs compared with lectures. That is, lecture course grades are typically weighted heavily towards high-stakes individual testing and exams, while lab grades typically rely on lower stakes assessments and group activities. Our study, however, provides additional nuance to this explanation: labs focused on concepts or skills both typically involve group work and we are unaware of systematic differences in testing strategies based on lab type. Future work should seek to identify additional types of activities that differ significantly between the three types of labs and with lectures to explain the remaining effects.

V. LIMITATIONS AND CONCLUSIONS

Our study is limited in several ways. First, our analysis focused exclusively on two physics lab assessments, which probe student views about experimental physics and critical thinking skills. We cannot say that the results can be generalized beyond these constructs or these assessments. The interpretations about students' views about experimental physics and critical thinking skills are only valid insofar as the assessments validly measure these constructs. Future work should further probe these ideas using additional data sources.

While the datasets are much larger and diverse than typical PER studies [22], simultaneously including data from an array of institution types and student characteristics, the analysis is still limited by sample size, in terms of both the number of unique classes and the number of students who identified with select demographic categories. The data are also not uniformly weighted between these variables, meaning some variables were more precisely measured than others and some results may be biased towards particular institutions or course types. The limited number of students in several demographic groups limited the reliability of estimates of the effect of lab type for those groups of students. Results presented in Appendix D 1 indicated that we could not have achieved much better precision for these estimates by using simpler models. Future work, therefore, should further test these results with more data with larger and more diverse samples.

Overall, we found that labs focused on skills improve (or produce equivalent) PLIC and E-CLASS scores for all students compared with labs aiming to reinforce concepts or do both, in part due to the increased focus on student decision-making and communication. The results have important implications for improving student learning and experiences in labs, as well as representation in physics. Given that students with more expertlike views tend to persist in physics [60], it is plausible that a focus on experimentation skills over reinforcing concepts (or, alternatively, providing increased focus on student decision making and communication in labs) could retain more women and students from backgrounds historically excluded and marginalized in physics. Future work should evaluate this possibility explicitly, acknowledging that the different types of labs may affect other aspects of student learning and experiences in different ways.

ACKNOWLEDGMENTS

We acknowledge support from NSF PHY-1734006 and NSF DUE-1611482.

APPENDIX A: DATA COLLECTION AND PROCESSING

Data were collected with the E-CLASS between August 2016 and December 2019 and data were collected with the PLIC between August 2017 and December 2020. Both the E-CLASS and PLIC were administered online as part of an automated administration system [56]. Individual instructors determined how to administer the instruments in their classes, but the automated system sent regular reminders to instructors updating them on how many of their students had completed the assessment. In Sec. A 1, we summarize

how data were filtered to arrive at the dataset used in the main text [61]. In Sec. A 2, we provide additional details on the student-level variables used in our analyses: gender, race or ethnicity, and major.

1. Data filtering

In total, 36 538 students in first-year classes submitted valid E-CLASS responses and 10 387 students in first-year classes submitted valid PLIC responses. We considered a response to the E-CLASS to be valid if the student

1. clicked submit at the end of the survey,
2. consented to participate in the study,
3. responded to at least one question, and
4. responded correctly to the filtering question used to eliminate responses from students who were not reading the questions.

We considered a response to the PLIC to be valid if the student:

1. clicked submit at the end of the survey,
2. consented to participate in the study,
3. indicated that they were at least 18 years of age, and
4. spent at least 30 sec on at least one of the four pages of the assessment.

For both assessments, if a student submitted more than one valid pre or post-test response, we kept only the first submitted valid response. Students also sometimes took the assessments multiple times as part of different classes. We treated these instances, denoted as student records in Table III, as independent events and used student records as the unit of analysis. We refer to student records as students in the main text.

We removed entire classes from our datasets when the instructor did not indicate the main purpose of their lab. We additionally removed classes from our datasets when the assessment was not administered both as a pretest and as a post-test. Both class-level filters were necessary to answer our research questions. We applied only one filter at the student level, removing students who did not have matched pre and post-test responses in the datasets. We matched students within classes by student ID for both the E-CLASS and PLIC. For the PLIC, we additionally matched students within classes by combination of first and last name. Students provided this information at the end of both assessments.

The student-level filtering aims primarily to improve the validity of the analysis by removing, for example, incomplete responses, responses from students who did not complete the course, or responses from individuals randomly clicking through the instrument. The filtering, however, may introduce systematic biases in the dataset, such as skewing towards students with higher grades and scores on the assessments [62]. The data collected from the PLIC between March 2020 and December 2020 are also likely biased based on who was able to participate due to the challenges of the COVID-19 pandemic. Imputation methods are recommended for mitigating such biases [63], but our data sources do not include sufficient information to accurately identify and impute the missing data.

Table III summarizes how the class and student-level filters affected our datasets. Our matched datasets for the E-CLASS and PLIC contained 16 409 and 4988 students, respectively. We did not have accurate information about how many students were enrolled in each class, but we calculated an upper bound on the response rates by assuming that all students enrolled in classes in our datasets completed *at least one* of the pre or post-test. With this assumption, 49.2% is an upper bound on the response rate for the E-CLASS and 52.4% is an upper bound on the response rate for the PLIC. These response rates are in line with typical response rates reported in the literature [63].

TABLE III. Number of institutions, classes, students, and student records included in the PLIC and E-CLASS datasets following each round of the data filtering process. We define a class as a combination of course (e.g., Physics 101) and semester (e.g., Fall 2019), so a single course may administer the E-CLASS or PLIC in multiple semesters and count as multiple classes. Students that took an assessment in different classes are counted multiple times in the student records tally. In the main text, we refer to student records as students and use these data in our analyses.

		Full dataset	Class filters	Student filter
E-CLASS	Institutions	65	59	56
	Classes	287	243	230
	Students	36 538	29 510	14 757
	Student records	41 514	33 384	16 409
PLIC	Institutions	34	28	28
	Classes	101	77	77
	Students	10 387	9024	4823
	Student records	11 577	9527	4988

2. Student-level variables

Students optionally provided demographic information at the end of the E-CLASS and PLIC. For the E-CLASS, students could select from three options when identifying their gender: woman, man, or other (with text box). For the PLIC, students could select from five options when identifying their gender: woman, man, nonbinary (with text box), prefer to self-describe (with text box), or "prefer not to disclose." We distinguished students' self-identified genders using the terms *man*, *woman*, and *nonbinary* (which included students that selected "other" on the E-CLASS or "nonbinary" or "prefer to self-describe" on the PLIC) to more closely align with nonbinary and fluid definitions of gender identity [64]. Students could select multiple racial or ethnic identities from a list of seven that we provided: *American Indian or Alaska Native*, *Asian*, *Black or African American*, *Hispanic or Latino*, *Native Hawaiian or other Pacific Islander*, *White*, or *other race or ethnicity*. Again, students could select prefer not to disclose or skip the question entirely. We did not treat these race or ethnicities as

TABLE IV. CIS items designed to measure the amount of decision making, modeling, and communication activities in labs. Items on the CIS asked instructors how often students engaged in the listed activities: *never*, *rarely*, *sometimes*, *often*, *always*. Two items (D1 and C4) were dropped from our revised measurement model for reasons discussed in the text. The factor loadings presented were calculated after standardizing the latent variables.

Factor	Code	Item	Loading
Decision making	D1	Develop their own research questions	...
	D2	Design their own procedures	1.149 ± 0.074
	D3	Build their own apparatus	0.939 ± 0.081
	D4	Choose their own analysis methods	0.919 ± 0.073
	D5	Troubleshoot problems with the setup or apparatus	0.703 ± 0.073
	D6	Refine system to reduce uncertainty	0.850 ± 0.081
Modeling	M1	Develop mathematical models for the system being studied	0.982 ± 0.096
	M2	Develop conceptual models for the system being studied	0.614 ± 0.096
	M3	Develop mathematical models for the measurement tools being used	0.534 ± 0.089
	M4	Develop conceptual models for the measurement tools being used	0.467 ± 0.084
	M5	Use mathematical or conceptual models to make predictions	0.468 ± 0.075
Communication	C1	Give oral presentations	0.512 ± 0.093
	C2	Write lab reports	0.458 ± 0.163
	C3	Maintain lab notebooks	0.989 ± 0.163
	C4	Read journal articles	...

mutually exclusive; rather, we included each of these race or ethnicities as separate independent variables in our analyses, so students could belong to multiple groups. The choices for students' race or ethnicity follows the Department of Education IPEDS definitions of race [65].

The E-CLASS and PLIC provided different options for students when selecting their intended major, but both assessments allowed students to select prefer not to disclose or to skip the question. We collapsed students' intended major on the E-CLASS into seven categories: *engineering*, *life science*, *math and computer science*, *physics* (including astronomy, astrophysics, and engineering physics), *other science* (including chemistry, geology, and geophysics), *nonscience*, and *open or undeclared*. The original version of the PLIC provided students with five options when selecting their intended major and we kept those original groups with one exception: we combined physics and engineering physics into one group, *physics*, consistent with our E-CLASS groups. We thus used four groups for students' intended major on the PLIC: *engineering*, *physics*, *other science*, and *other*. We, again, labeled a student's major as *unknown* when this information was not provided by the student.

APPENDIX B: MEASUREMENT MODEL FOR PEDAGOGICAL FEATURES OF LABS

In a previous analysis using a similar dataset, Holmes and Lewandowski identified (using both exploratory and confirmatory factor analysis) a group of items on the CIS that measured the amount of decision-making and modeling in labs [35]. We began with these items and factor structure when constructing a measurement model for the amount of decision-making and modeling in labs during our analysis. We extended the model examined in Ref. [35] by including an additional factor for the amount of communication activities in labs using additional items from the CIS. All of the CIS items used in this analysis are listed in Table IV.

We first performed a confirmatory factor analysis using the three-factor model presented in Table IV. We found that this measurement model did not adequately describe the data (confirmatory fit index $[\mathrm{CFI}] = 0.793$; root mean square error of approximation $[\mathrm{RMSEA}] = 0.134$; standardized root mean square residual $[\mathrm{SRMR}] = 0.100$). Examining the standardized factor loadings, we found that item C4 did not load strongly onto the hypothesized communication factor (standardized factor loading <0.35). We also found that item D1 had large residual correlations with three other items, including two that were part of the hypothesized communication factor (which was not part of the original model developed in Ref. [8]).

In our revised model, we removed items D1 and C4. We also added covariance terms between modeling items with parallel language (i.e., M1 and M2, M1 and M3, M2 and M4, and M3 and M4) and with large modification indices in the original model (the average modification index for these four terms was 23.5). We found that this revised model fit our data adequately ($\mathrm{CFI} = 0.909$; $\mathrm{RMSEA} = 0.100$; $\mathrm{SRMR} = 0.074$). The factor loadings from this model are shown in Table IV.

APPENDIX C: FULL RESULTS OF LINEAR MIXED MODELS

1. RQ1: Effects of different lab types on scores for different subpopulations of students

We present complete results of fitted models from Sec. III A in Table V. Standardized effect sizes were obtained by standardizing continuous variables (i.e., pretest

TABLE V. Two-level linear mixed models of E-CLASS and PLIC post-test scores including lab type and demographic variables with institutions as random intercepts. Standardized coefficients represent effects with pretest and post-test scores grand mean centered. Effects for nominal variables are relative to the reference level (variable level not shown). The variance explained by the fixed effects and random effects (fixed effects alone) is 44% (42%) for E-CLASS and 16% (10%) for PLIC.

	Dependent variable							
	E-CLASS post-test score				PLIC post-test score			
	β	SE	β_{std}	SE	β	SE	β_{std}	SE
Pretest	0.716***	0.007	0.614***	0.006	0.291***	0.014	0.273***	0.013
Lab type								
Mixed	0.687	0.515	0.091	0.066	−0.296	0.186	−0.254	0.155
Skills based	1.046	0.616	0.138	0.079	0.160	0.141	0.137	0.117
Gender								
Nonbinary	−2.327*	1.088	−0.307*	0.140	0.331	0.332	0.283	0.277
Woman	−1.073***	0.210	−0.141***	0.027	0.026	0.053	0.022	0.044
Unknown	−2.116**	0.730	−0.279**	0.094	0.216	0.386	0.185	0.322
Race or ethnicity								
American Indian	−0.867	1.390	−0.114	0.178	0.173	0.235	0.148	0.196
Asian	0.337	0.420	0.044	0.054	−0.093	0.101	−0.080	0.084
Black	−0.277	0.491	−0.036	0.063	−0.266	0.171	−0.228	0.142
Hispanic	−0.423	0.444	−0.056	0.057	−0.069	0.105	−0.059	0.088
Native Hawaiian	−1.634	1.151	−0.215	0.148	0.127	0.347	0.109	0.290
White	−0.211	0.396	−0.028	0.051	0.024	0.099	0.021	0.082
Other	−1.278	0.745	−0.168	0.096	−0.363*	0.157	−0.311*	0.131
Unknown	0.023	0.626	0.003	0.080	−0.106	0.202	−0.090	0.169
Major								
Engineering	−0.939***	0.173	−0.124***	0.022	−0.189***	0.052	−0.162***	0.043
Life Sciences	−1.647***	0.196	−0.217***	0.025				
Math or CS	−1.380***	0.194	−0.182***	0.025				
Other science	−1.483***	0.200	−0.196***	0.026	−0.155**	0.052	−0.133**	0.043
Nonscience	−2.431***	0.231	−0.320***	0.030				
Undeclared	−0.584	0.318	−0.077	0.041				
Other					−0.292***	0.067	−0.250***	0.056
Unknown	−1.707*	0.728	−0.225*	0.093	−0.229*	0.109	−0.196*	0.091
Lab type * Gender								
Mixed * Nonbinary	0.749	1.240	0.099	0.159	−0.900	0.472	−0.770	0.393
Skills based * Nonbinary	1.486	1.375	0.196	0.177	−0.597	0.389	−0.511	0.325
Mixed * Woman	0.777**	0.243	0.102**	0.031	0.189*	0.095	0.161*	0.079
Skills based * Woman	1.264***	0.288	0.167***	0.037	−0.041	0.072	−0.035	0.060
Mixed * Unknown	2.042*	0.841	0.269*	0.108	−0.611	0.582	−0.523	0.485
Skills based * Unknown	1.185	1.075	0.156	0.138	0.033	0.478	0.028	0.399
Lab type * Race or ethnicity								
Mixed * American Indian	0.559	1.492	0.074	0.192	−0.532	0.399	−0.455	0.333
Skills based * American Indian	1.854	1.856	0.244	0.238	−0.586	0.311	−0.501	0.259
Mixed * Asian	−0.064	0.476	−0.008	0.061	0.147	0.172	0.126	0.143
Skills based * Asian	−0.362	0.581	−0.048	0.075	0.024	0.132	0.021	0.110
Mixed * Black	−0.844	0.590	−0.111	0.076	0.197	0.257	0.169	0.215
Skills based * Black	0.408	0.802	0.054	0.103	0.092	0.206	0.078	0.171
Mixed * Hispanic	−0.079	0.515	−0.010	0.066	0.255	0.185	0.219	0.154
Skills based * Hispanic	0.501	0.615	0.066	0.079	0.059	0.136	0.051	0.113
Mixed * Native Hawaiian	1.302	1.289	0.172	0.166	0.464	0.543	0.397	0.453
Skills based * Native Hawaiian	2.384	1.559	0.314	0.200	−0.389	0.519	−0.333	0.432
Mixed * White	0.280	0.452	0.037	0.058	0.364*	0.168	0.312*	0.140
Skills based * White	0.313	0.558	0.041	0.072	0.108	0.128	0.092	0.107
Mixed * Other	0.516	0.847	0.068	0.109	0.254	0.229	0.217	0.191

(Table continued)

TABLE V. *(Continued)*

	Dependent variable							
	E-CLASS post-test score				PLIC post-test score			
	β	SE	β_{std}	SE	β	SE	β_{std}	SE
Skills based * Other	1.177	1.021	0.155	0.131	0.368	0.195	0.315	0.163
Mixed * Unknown	−0.095	0.714	−0.012	0.092	0.216	0.334	0.185	0.278
Skills based * Unknown	−0.740	0.841	−0.097	0.108	0.094	0.266	0.081	0.221
Constant	4.621***	0.513	0.140*	0.064	3.844***	0.154	−0.088	0.111

*$p < 0.05$.
**$p < 0.01$.
***$p < 0.001$.

TABLE VI. Two-level linear mixed models of E-CLASS post-test scores including pedagogical and demographic variables with institution as a random intercept. Standardized coefficients represent effects with pretest and post-test scores grand mean centered. For nominal variables, effects are estimated relative to the reference level, which can be inferred by the level of the variable that is not shown in the table. The percent variance explained by the fixed effects alone is approximately 41% for both models. The percent variance explained by the fixed and random effects together is approximately 44% for both models.

	Dependent variable: E-CLASS post-test score							
	Model 1				Model 2			
	β	SE	β_{std}	SE	β	SE	β_{std}	SE
Pretest	0.715***	0.007	0.613***	0.006	0.715***	0.007	0.612***	0.006
Pedagogical variables								
Decision making	0.455***	0.096	0.056***	0.012				
Modeling	0.285**	0.095	0.034**	0.011	0.298***	0.090	0.035***	0.010
Communication					0.500***	0.088	0.061***	0.010
Gender								
Nonbinary	−1.449**	0.443	−0.191***	0.057	−1.454**	0.443	−0.192***	0.057
Woman	−0.337***	0.097	−0.044***	0.012	−0.334***	0.097	−0.044***	0.012
Unknown	−0.603	0.336	−0.079	0.043	−0.591	0.336	−0.078	0.043
Race or ethnicity								
American Indian	−0.225	0.467	−0.030	0.060	−0.217	0.466	−0.029	0.060
Asian	0.228	0.179	0.030	0.023	0.222	0.179	0.029	0.023
Black	−0.683**	0.248	−0.090**	0.032	−0.675**	0.248	−0.089**	0.032
Hispanic	−0.308	0.200	−0.041	0.026	−0.309	0.199	−0.041	0.026
Native Hawaiian	−0.305	0.465	−0.040	0.060	−0.298	0.465	−0.039	0.060
White	0.021	0.172	0.003	0.022	0.020	0.172	0.003	0.022
Other	−0.722*	0.315	−0.095*	0.040	−0.721*	0.315	−0.095*	0.040
Unknown	−0.182	0.265	−0.024	0.034	−0.193	0.265	−0.025	0.034
Major								
Engineering	−0.947***	0.173	−0.125***	0.022	−0.929***	0.173	−0.122***	0.022
Life Sciences	−1.661***	0.196	−0.219***	0.025	−1.635***	0.196	−0.216***	0.025
Math or CS	−1.386***	0.195	−0.183***	0.025	−1.373***	0.194	−0.181***	0.025
Other science	−1.500***	0.201	−0.198***	0.026	−1.479***	0.201	−0.195***	0.026
Nonscience	−2.422***	0.231	−0.319***	0.030	−2.414***	0.231	−0.318***	0.030
Undeclared	−0.589	0.319	−0.078	0.041	−0.605	0.318	−0.080	0.041
Unknown	−1.680*	0.728	−0.221*	0.093	−1.683*	0.728	−0.222*	0.093
Constant	5.136***	0.331	0.170***	0.040	5.093***	0.326	0.166***	0.039

*$p < 0.05$.
**$p < 0.01$.
***$p < 0.001$.

and post-test scores) and refitting the models. In the main text, we reported expected outcomes from our fitted models (i.e., marginal effects), which indicate how post-test scores varied with particular independent variables (i.e., lab type, gender, and race or ethnicity).

2. RQ3: Effects of pedagogical variables on students' scores

We present complete results of fitted models from Sec. III C in Tables VI and VII. Standardized effect sizes were obtained by standardizing continuous variables (i.e., pretest and post-test scores) and refitting the models. In the main text, we reported expected outcomes from our fitted models (i.e., marginal effects), which indicate how post-test scores varied with particular independent variables (i.e., pedagogical features).

In Sec. III C, we also reported estimates for the fraction of the effect of lab type that could be explained by pedagogical features of the labs. We calculated these estimates by combining the results of Secs. III A and III B with the results presented in Tables VI and VII above. In Sec. III B, we calculated the differences in average factor scores between skills-based and concepts-based labs ($\mathbb{E}[\mathrm{FS}_{\mathrm{diff}}]$ in Table VIII). Multiplying these difference by the standardized effects of the pedagogical variables in Tables VI and VII (β_{std}), we obtained an estimate for the expected difference in post-test scores (in units of standard deviations) between skills-based labs and concepts-based labs, considering only the pedagogical variables. Dividing this value by the marginal effect of skills-based labs (compared to concepts-based labs) on students' standardized post-test scores from Sec. III A (0.21 ± 0.05 for the E-CLASS and 0.18 ± 0.10 for the PLIC), we obtained an estimate of the fraction of the marginal

TABLE VII. Two-level linear mixed models of PLIC post-test scores including pedagogical and demographic variables with institution as a random intercept. Standardized coefficients represent effects with pretest and post-test scores grand mean centered. For nominal variables, effects are estimated relative to the reference level, which can be inferred by the level of the variable that is not shown in the table. The percent variance explained by the fixed effects alone is approximately 10% for both models. The percent variance explained by the fixed and random effects together is approximately 16% for model 1 and approximately 17% for model 2.

	Dependent variable: PLIC post-test score							
	Model 1				Model 2			
	β	SE	β_{std}	SE	β	SE	β_{std}	SE
Pretest	0.294***	0.015	0.275***	0.013	0.295***	0.015	0.276***	0.013
Pedagogical variables								
Decision making	0.085***	0.025	0.078***	0.022				
Modeling	0.024	0.043	0.019	0.032	0.044	0.038	0.035	0.029
Communication					0.097***	0.022	0.093***	0.021
Gender								
Nonbinary	−0.193	0.154	−0.165	0.128	−0.196	0.154	−0.167	0.128
Woman	0.031	0.033	0.026	0.028	0.031	0.033	0.027	0.027
Unknown	0.189	0.211	0.161	0.175	0.183	0.210	0.156	0.175
Race or ethnicity								
American Indian	−0.218	0.139	−0.186	0.116	−0.217	0.139	−0.185	0.116
Asian	−0.060	0.060	−0.051	0.050	−0.061	0.060	−0.052	0.050
Black	−0.176*	0.086	−0.150*	0.072	−0.183*	0.086	−0.156*	0.072
Hispanic	−0.003	0.062	−0.003	0.051	−0.005	0.062	−0.0004	0.051
Native Hawaiian	0.120	0.219	0.102	0.182	0.128	0.219	0.109	0.182
White	0.123*	0.058	0.105*	0.048	0.123*	0.058	0.105*	0.048
Other	−0.152	0.083	−0.129	0.069	−0.151	0.083	−0.129	0.069
Unknown	−0.034	0.120	−0.029	0.100	−0.034	0.120	−0.029	0.099
Major								
Engineering	−0.209***	0.052	−0.178***	0.044	−0.195***	0.053	−0.167***	0.044
Other science	−0.181***	0.053	−0.155***	0.044	−0.173***	0.052	−0.148***	0.044
Other	−0.304***	0.068	−0.259***	0.057	−0.298***	0.068	−0.254***	0.057
Unknown	−0.232*	0.111	−0.198*	0.092	−0.224*	0.111	−0.191*	0.092
Constant	3.897***	0.123	−0.052	0.083	3.880***	0.123	−0.068	0.083

*$p < 0.05$.
**$p < 0.01$.
***$p < 0.001$.

TABLE VIII. Estimate of the fraction of the observed effect of skills-based labs (compared to concepts-based labs) that can be explained by the pedagogical variables. $\mathbb{E}[\mathrm{FS}_{\mathrm{diff}}]$ is the difference in average factor scores for skills-based and concepts-based labs. β_{std} is the effect of the pedagogical variable on students' post-test scores. $\mathbb{E}[\mathrm{FS}_{\mathrm{diff}}] \times \beta_{\mathrm{std}}$ gives the expected difference in students' post-test scores between skills-based and concepts-based labs with average factor scores, and dividing this value by the marginal effect of skills-based labs (compared to concepts-based labs), $\mathrm{ME}_{\mathrm{Skills}}$, gives an estimate of the proportion of the effect of skills-based labs that can be attributed to each pedagogical variable. Note that the decision-making and communication variables were modeled separately and so the effects of these variables are not additive; there is considerable overlap in the variance explained by these variables.

Variable	$\mathbb{E}[\mathrm{FS}_{\mathrm{diff}}]$	β_{std}	$\mathbb{E}[\mathrm{FS}_{\mathrm{diff}}] \times \beta_{\mathrm{std}}$	$\frac{\mathbb{E}[\mathrm{FS}_{\mathrm{diff}}] \times \beta}{\mathrm{ME}_{\mathrm{Skills}}}$
		E-CLASS		
Decision making	1.44 ± 0.17	0.056 ± 0.012	0.080 ± 0.019	0.342 ± 0.102
Modeling	0.45 ± 0.23	0.034 ± 0.011	0.015 ± 0.009	0.065 ± 0.041
Communication	1.57 ± 0.17	0.061 ± 0.010	0.096 ± 0.019	0.407 ± 0.111
		PLIC		
Decision making	1.44 ± 0.17	0.078 ± 0.022	0.112 ± 0.034	0.584 ± 0.329
Modeling	0.45 ± 0.23	0.019 ± 0.032	0.009 ± 0.015	0.045 ± 0.083
Communication	1.57 ± 0.17	0.093 ± 0.021	0.145 ± 0.036	0.756 ± 0.403

effect of lab type that can be explained by each of the pedagogical variables. These results are presented in Table VIII.

APPENDIX D: MODEL DIAGNOSTICS

In this appendix, we examine qualities of our linear mixed models from Sec. III A. In Sec. D 1 we examine how our model choices impacted precision in our estimated effects, while in Sec. D 2 we check visually how well our models satisfied assumptions of linear mixed models and discuss implications of violations of these assumptions.

1. Variance inflation factors

We prioritized accuracy over precision in this study by simultaneously controlling for several variables to better estimate the effect of lab type on students' scores. In this section, we present variance inflation factors that quantify the degree to which we decreased precision in our estimates by taking this approach. Variance inflation can have significant impacts on p values and elevate false negative rates, which is why we expressly avoided placing significant weight on p values in our discussion.

Variance inflation factors (VIFs) can be interpreted as the ratio of the standard error of a coefficient in a model to the standard error of that coefficient if only that variable, and no others, were included in the model. A VIF of two would indicate that the standard error of a coefficient in the model was double to its standard error in a model with only that variable. We used a generalized VIF (GVIF) here that corrects for the degrees of freedom of a variable [66] and is more useful for linear mixed models. We checked the GVIFs for all variables included in each of our linear mixed models. The results are shown in Table IX.

The GVIFs for most of the main effects variables were above 2, suggesting limited precision on the estimates of those effects. Our analysis was particularly concerned with the interaction terms between lab type, gender, and race or ethnicity, and the large VIFs on the main effects do not suggest any issues with the interpretation of the interaction terms. Models with interaction terms are generally

TABLE IX. Generalized variance inflation factors corrected for degrees of freedom ($\mathrm{GVIF}^{1/2df}$) for linear mixed models presented in the main text. Values roughly indicate the ratio of the standard error for a variable relative to the standard error for a model with only that variable included.

	E-CLASS	PLIC
Lab type	2.772	2.788
Pretest score	1.016	1.010
Major	1.010	1.025
Gender	3.376	1.963
American Indian	2.992	1.713
Asian	3.829	2.889
Black	2.301	2.322
Hispanic	2.787	1.949
Native Hawaiian	2.488	1.600
White	4.122	2.966
Other	2.604	1.983
Unknown Race	3.399	2.155
Lab type * Gender	1.593	1.482
Lab type * American Indian	1.741	1.323
Lab type * Asian	2.647	2.248
Lab type * Black	1.568	1.674
Lab type * Hispanic	1.976	1.534
Lab type * Native Hawaiian	1.585	1.271
Lab type * White	3.400	2.771
Lab type * Other	1.716	1.456
Lab type * Unknown race	2.240	1.647

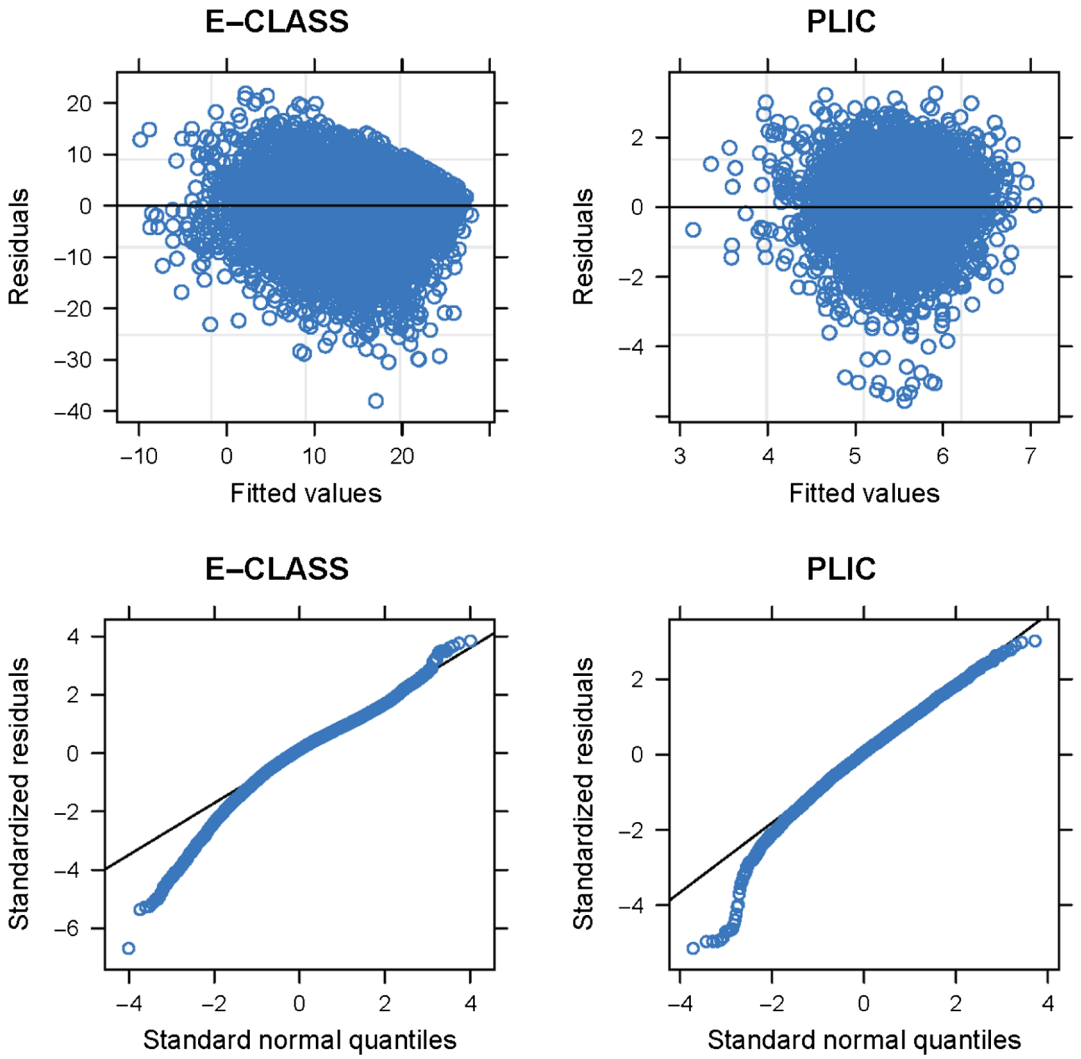

FIG. 4. Model diagnostic plots for the linear mixed models. There is no noticeable trend in the residuals for either model. There are generally departures from normality in the residuals at the tails of the distribution, however, which is not uncommon and does not generally affect the interpretation of p values [67].

susceptible to inflated variances, but we found only minimal variance inflation in the interaction terms here. Only the interaction terms for lab type with *Asian*, *White*, and *Unknown race* had GVIFs larger than two. The small GVIFs for the other interaction terms indicated that we could not have achieved substantially better precision for the interaction terms even if we had used a simpler model. The larger GVIFs on the interaction terms were not problematic for interpreting our data because, even with inflated variances, we were able to measure these terms more precisely than the others.

2. Visual check of model assumptions

Linear mixed models have the same modeling assumptions as multiple linear regression, in addition to assuming that there exists a nested structure to the data. The two most important assumptions that we evaluate here are the homoskedasticity of residuals with fitted values (i.e., the spread of residuals is approximately the same for all predicted values of the dependent variable) and the normality of residuals.

We evaluated the above assumptions visually using Fig. 4. Plots of residuals against fitted values did not display any obvious trends that would lead us to conclude that the assumption of homoskedasticity was violated egregiously. All quantile-quantile plots of standardized residuals displayed some departure from normality at the left tails of the distribution. This departure from normality at the tails of the distribution is not uncommon and does not generally affect the interpretation of p values or the estimated effects [67].

[1] Steve Olson and Donna Gerardi Riordan, *Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics*, Report to the president (Executive Office of the President, Washington, DC, 2012), https://aaas-arise.org/resource/engage-to-excel-producing-one-million-additional-college-graduates-with-degrees-in-science-technology-engineering-and-mathematics/.

[2] National Research Council *et al.*, *America's Lab Report: Investigations in High School Science* (National Academies Press, Washington, DC, 2006), https://www.nap.edu/catalog/11311/americas-lab-report-investigations-in-high-school-science.

[3] Joseph Kozminski, Heather Lewandowski, Nancy Beverly, Steve Lindaas, Duane Deardorff, Ann Reagan, Richard Dietz, Randy Tagg, M. EblenZayas, J. Williams *et al.*, *AAPT Recommendations for the Undergraduate Physics Laboratory Curriculum*, (American Association of Physics Teachers, USA, 2014), https://www.aapt.org/resources/upload/labguidlinesdocument_ebendorsed_nov10.pdf.

[4] Ricardo Trumper, The physics laboratory–a historical overview and future perspectives, Sci. Educ. **12**, 645 (2003).

[5] Robin Millar *et al.*, *The Role of Practical Work in the Teaching and Learning of Science* (Committee on High School Science Laboratories: Role and Vision; National Research Council; National Academy of Sciences, Washington DC, 2004), p. 308, https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_073330.pdf.

[6] Benjamin M. Zwickl, Noah Finkelstein, and Heather J. Lewandowski, The process of transforming an advanced lab course: Goals, curriculum, and assessments, Am. J. Phys. **81**, 63 (2013).

[7] Avi Hofstein and Vincent N. Lunetta, The laboratory in science education: Foundations for the twenty-first century, Sci. Educ. **88**, 28 (2004).

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

[9] National Research Council, *Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering*, Tech. Rep. (Committee on the Status, Contributions, and Future Directions of Discipline-Based Education Research; Board on Science Education; Division of Behavioral and Social Sciences and Education; National Research Council, Washington, DC, 2012), https://www.nap.edu/catalog/13362/discipline-based-education-research-understanding-and-improving-learning-in-undergraduate.

[10] Sandra Laursen, *Levers for Change: An Assessment of Progress on Changing STEM Instruction*, Tech. Rep. (American Association for the Advancement of Science, Washington, DC, 2019), https://www.aaas.org/resources/levers-change-assessment-progress-changing-stem-instruction.

[11] Marie-Geneviève Séré, Towards renewed research questions from the outcomes of the European project labwork in science education, Sci. Educ. **86**, 624 (2002).

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

[13] N. G. Holmes, Jack Olsen, James L. Thomas, and Carl E. Wieman, Value added or misattributed? A multi-institution study on the educational benefit of labs for reinforcing physics content, Phys. Rev. Phys. Educ. Res. **13**, 010129 (2017).

[14] Emily M. Smith, Martin M. Stein, Cole Walsh, and N. G. Holmes, Direct Measurement of the Impact of Teaching Experimentation in Physics Labs, Phys. Rev. X **10**, 011029 (2020).

[15] Eugenia Etkina, Anna Karelina, Maria Ruibal-Villasenor, David Rosengrant, Rebecca Jordan, and Cindy E. Hmelo-Silver, Design and reflection help students develop scientific abilities: Learning in introductory physics laboratories, J. Learn. Sci. **19**, 54 (2010).

[16] Eugenia Etkina, Alan Van Heuvelen, Anna Karelina, Maria Ruibal-Villasenor, David Rosengrant, Leon Hsu, Charles Henderson, and Laura McCullough, Spending time on design: Does it hurt physics learning?, AIP Conf. Proc. **951**, 88 (2007).

[17] Bethany R. Wilcox and H. J. Lewandowski, Developing skills versus reinforcing concepts in physics labs: Insight from a survey of students' beliefs about experimental physics, Phys. Rev. Phys. Educ. Res. **13**, 010108 (2017).

[18] Rachel Henderson, Kelsey Funkhouser, and Marcos Caballero, A longitudinal exploration of students' beliefs about experimental physics, in *Proceedings of the 2019 Physics Education Research Conference, Provo, UT* (AIP, New York, 2019).

[19] N. G. Holmes, Carl E. Wieman, and D. A. Bonn, Teaching critical thinking, Proc. Natl. Acad. Sci. U.S.A. **112**, 11199 (2015).

[20] Cole Walsh, Katherine N. Quinn, and Natasha Holmes, Assessment of critical thinking in physics labs: Concurrent validity, in *Proceedings of the 2018 Physics Education Research Conference, Washington, DC* (AIP, New York, 2018).

[21] Cole Walsh, Katherine N. Quinn, C. Wieman, and N. G. Holmes, Quantifying critical thinking: Development and validation of the physics lab inventory of critical thinking, Phys. Rev. Phys. Educ. Res. **15**, 010135 (2019).

[22] Stephen Kanim and Ximena C. Cid, Demographics of physics education research, Phys. Rev. Phys. Educ. Res. **16**, 020106 (2020).

[23] Adrian Madsen, Sarah B. McKagan, 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] Ben Van Dusen and Jayson Nissen, Equity in college physics student learning: A critical quantitative intersectionality investigation, J. Res. Sci. Teach. **57**, 33 (2020).

[25] Bethany 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).

[26] James Day, Jared B. Stang, N. G. Holmes, Dhaneesh Kumar, and D. A. Bonn, Gender gaps and gendered action

in a first-year physics laboratory, Phys. Rev. Phys. Educ. Res. **12**, 020104 (2016).

[27] Rebecca L. Matz, Benjamin P. Koester, Stefano Fiorini, Galina Grom, Linda Shepard, Charles G. Stangor, Brad Weiner, and Timothy A. McKay, Patterns of gendered performance differences in large introductory courses at five research universities, AERA Open **3**, 1 (2017).

[28] Jasna Jovanovic and Sally Steinbach King, Boys and girls in the performance-based science classroom: Who's doing the performing?, Am. Educ. Res. J. **35**, 477 (1998).

[29] N. G. Holmes, Ido Roll, and D. A. Bonn, Participating in the physics lab: Does gender matter?, Phys. Canada **40**, 84 (2014).

[30] 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).

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

[32] Anna Teresia Danielsson and Cedric Linder, Learning in physics by doing laboratory work: Towards a new conceptual framework, Gender Educ. **21**, 129 (2009).

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

[34] Andrew Elby and David Hammer, On the substance of a sophisticated epistemology, Sci. Educ. **85**, 554 (2001).

[35] N. G. Holmes, Benjamin Keep, and Carl E. Wieman, Developing scientific decision making by structuring and supporting student agency, Phys. Rev. Phys. Educ. Res. **16**, 010109 (2020).

[36] Jennifer A. Schmidt, Joshua M. Rosenberg, and Patrick N. Beymer, A person-in-context approach to student engagement in science: Examining learning activities and choice, J. Res. Sci. Teach. **55**, 19 (2018).

[37] 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).

[38] Angela Calabrese Barton and Edna Tan, We be burnin'! Agency, identity, and science learning, J. Learn. Sci. **19**, 187 (2010).

[39] Malcolm Wells, David Hestenes, and Gregg Swackhamer, A modeling method for high school physics instruction, Am. J. Phys. **63**, 606 (1995).

[40] Eric Brewe, Modeling theory applied: Modeling instruction in introductory physics, Am. J. Phys. **76**, 1155 (2008).

[41] Benjamin M. Zwickl, Dehui Hu, Noah Finkelstein, and H. J. Lewandowski, Model-based reasoning in the physics laboratory: Framework and initial results, Phys. Rev. ST Phys. Educ. Res. **11**, 020113 (2015).

[42] Jacob T. Stanley, Weifeng Su, and H. J. Lewandowski, Using lab notebooks to examine students' engagement in modeling in an upper-division electronics lab course, Phys. Rev. Phys. Educ. Res. **13**, 020127 (2017).

[43] Dimitri R. Dounas-Frazer, Kevin L. Van De Bogart, MacKenzie R. Stetzer, and H. J. Lewandowski, Investigating the role of model-based reasoning while troubleshooting an electric circuit, Phys. Rev. Phys. Educ. Res. **12**, 010137 (2016).

[44] Eric Brewe, Laird Kramer, and George O'Brien, Modeling instruction: Positive attitudinal shifts in introductory physics measured with CLASS, Phys. Rev. Phys. ST Educ. Res. **5**, 013102 (2009).

[45] Ronald L Wasserstein and Nicole A Lazar, The ASA statement on *p* values: Context, process, and purpose, The American Statistician **70**, 129 (2016).

[46] 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).

[47] Jessica R. Hoehn and H. J. Lewandowski, Incorporating writing in advanced lab projects: A multiple case-study analysis, Phys. Rev. Phys. Educ. Res. **16**, 020161 (2020).

[48] Paula Heron, Phys21: Preparing physics students for 21st century careers, Bull. Am. Phys. Soc. **62** (2017), https://www.compadre.org/jtupp/report.cfm.

[49] Karyn L. Lewis, Jane G. Stout, Steven J. Pollock, Noah D. Finkelstein, and Tiffany A. Ito, Fitting in or opting out: A review of key social-psychological factors influencing a sense of belonging for women in physics, Phys. Rev. Phys. Educ. Res. **12**, 020110 (2016).

[50] Bethany R. Wilcox and Heather 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).

[51] Eric Burkholder, Cole Walsh, and N. G. Holmes, Examination of quantitative methods for analyzing data from concept inventories, Phys. Rev. Phys. Educ. Res. **16**, 010141 (2020).

[52] Idaykis Rodriguez, Eric Brewe, Vashti Sawtelle, and Laird H. Kramer, Impact of equity models and statistical measures on interpretations of educational reform, Phys. Rev. ST Phys. Educ. Res. **8**, 020103 (2012).

[53] Ben Van Dusen and Jayson Nissen, Associations between learning assistants, passing introductory physics, and equity: A quantitative critical race theory investigation, Phys. Rev. Phys. Educ. Res. **16**, 010117 (2020).

[54] Benjamin M. Zwickl, Takako Hirokawa, Noah Finkelstein, and Heather J. Lewandowski, Epistemology and expectations survey about experimental physics: Development and initial results, Phys. Rev. ST Phys. Educ. Res. **10**, 010120 (2014).

[55] Cole Walsh and Natasha Holmes, Assessing the assessment: Mutual information between response choices and factor scores, in *Proceedings of the 2019 Physics Education Research Conference, Provo, UT* (AIP, New York, 2019).

[56] 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).

[57] Devyn Shafer, Maggie S. Mahmood, and Tim Stelzer, Impact of broad categorization on statistical results: How underrepresented minority designation can mask the struggles of both Asian American and African American students, Phys. Rev. Phys. Educ. Res. **17**, 010113 (2021).

[58] Ben Van Dusen and Jayson Nissen, Modernizing use of regression models in physics education research: A review of hierarchical linear modeling, Phys. Rev. Phys. Educ. Res. **15**, 020108 (2019).

[59] L. L. Thurstone, *The Vectors of Mind* (University of Chicago Press, Chicago, IL, 1935).

[60] Katherine K. Perkins and Mindy Gratny, Who becomes a physics major? A long-term longitudinal study examining the roles of pre-college beliefs about physics and learning physics, interest, and academic achievement, AIP Conf. Proc. **1289**, 253 (2010).

[61] John M. Aiken, Riccardo De Bin, H. J. Lewandowski, and Marcos D. Caballero, Framework for evaluating statistical models in physics education research, Phys. Rev. Phys. Educ. Res. **17**, 020104 (2021).

[62] Jayson M. Nissen, Manher Jariwala, Eleanor W. Close, and Ben Van Dusen, Participation and performance on paper- and computer-based low-stakes assessments, Int. J. STEM Educ. **5**, 21 (2018).

[63] Jayson Nissen, Robin Donatello, and Ben Van Dusen, Missing data and bias in physics education research: A case for using multiple imputation, Phys. Rev. Phys. Educ. Res. **15**, 020106 (2019).

[64] 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).

[65] IPEDS, IPEDS definitions, Accessed: 10-23-2018, https://nces.ed.gov/ipeds/report-your-data/race-ethnicity-reporting-changes.

[66] John Fox and Georges Monette, Generalized collinearity diagnostics, J. Am. Stat. Assoc. **87**, 178 (1992).

[67] Thomas Lumley, Paula Diehr, Scott Emerson, and Lu Chen, The importance of the normality assumption in large public health data sets, Annual review of public health **23**, 151 (2002).