

Group dynamics in inquiry-based labs: Gender inequities and the efficacy of partner agreements

Matthew Dew¹,[✉] Emma Hunt²,[✉] Viranga Perera²,[✉] Jonathan Perry²,[✉]
 Gregorio Ponti³,[✉] and Andrew Loveridge^{2,*}

¹*Laboratory of Atomic and Solid State Physics, Cornell University, Ithaca, New York 14853, USA*

²*Department of Physics, The University of Texas at Austin, Austin, Texas 78712, USA*

³*Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA*



(Received 30 April 2023; accepted 12 January 2024; published 5 April 2024)

[This paper is part of the Focused Collection on Instructional labs: Improving traditions and new directions.] Recent studies provide evidence that social constructivist pedagogical methods such as active learning, interactive engagement, and inquiry-based learning, while pedagogically more effective, can enable inequities in the classroom. By conducting a quantitative empirical examination of gender-inequitable group dynamics in two inquiry-based physics labs, we extend results of previous work. Using a survey on group work preferences and video recordings of lab sessions, we find similar patterns of gendered role taking noted in prior studies. These results are not reducible to differences in students' preferences. We find that an intervention which employed partner agreement forms, with the goal of reducing inequities, had a positive impact on students' engagement with equipment during a first-semester lab course. Our work will inform implementation of more effective interventions in the future and emphasizes challenges faced by instructors who are dedicated to both research-based pedagogical practices and efforts to promote diversity, equity, and inclusion in their classrooms.

DOI: [10.1103/PhysRevPhysEducRes.20.010121](https://doi.org/10.1103/PhysRevPhysEducRes.20.010121)

I. INTRODUCTION

Group work is a common feature of many university physics lecture and lab courses. Research-based pedagogical practices, like active learning, interactive engagement, and inquiry-based learning, employ varying degrees of group work. Beyond its role as a pedagogical tool, group work has been identified as a learning goal itself. The American Association of Physics Teachers has designated it a component of scientific collaboration and as a learning goal for lab courses in particular [1]. However, despite its potential benefits, group work introduces complexities into a course which can produce unintended effects.

Pedagogical practices that incorporate group work such as active learning, interactive engagement, and inquiry-based learning enjoy broad support within the physics education community. This support is based on a breadth of empirical studies that show them to be pedagogically more effective than traditional lecture or lab courses [2,3]. However, while these practices may be best for learning overall, there is also evidence that they can enable certain inequities. For example, Quinn *et al.* [4] observed that

incorporating inquiry-based instructional practices into laboratory courses, compared directly with traditional labs, can result in an increase of gendered role taking. Other studies have found this gendered division of labor to women's disadvantage [5–7]. In the context of lecture courses, Gordon *et al.* [8] found that a flipped classroom had a negative impact on learning and achievement for low-income, systemically nondominant race or ethnicity,¹ and first-generation students when compared with an interactive lecture. These recent works corroborate the results of other studies, which show more generally that—absent proactive efforts by an instructor—systemically nondominant groups engage less in active-learning components of lecture courses [10–15]. Collectively, these empirical observations suggest that many popular research-based pedagogical methods can be especially susceptible to unintended inequitable outcomes. As we will review in Sec. II, there are also sound theoretical reasons to expect this. This situation presents instructors with a serious tension: pedagogical methods that research suggests are best for overall student learning can be worse for diversity, equity, and inclusion.

Of course, this is by no means inevitable. In some contexts, research-based remediation strategies have been developed and successfully implemented [11,12]. It is an

*Andrew.Loveridge@austin.utexas.edu

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/). Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

¹We use the term “systemically nondominant” as explained in [9].

important goal to build upon this work, especially to include methods designed specifically for inquiry-based lab courses and lab courses more generally.

Given the integral role of group work in these pedagogical methods, it should not be surprising that one major source of these inequities lies in problematic group dynamics. The value of group work, its potential for inequities, and frameworks for promoting fair and effective group work have been important topics in physics education research for several decades, dating back to a multiyear study at the University of Minnesota [16,17]. More recent research has explored this issue in depth in the context of laboratory courses, finding cross-cultural evidence for gendered division of labor [18], documenting the likelihood of women to adopt secretary or project manager roles [7,19], and assessing how the frequency of intergroup interactions is affected by lab design and group gender composition [20]. However, these studies are limited to three institutions and it remains unclear what strategies can be used to resolve the inequities they identify.

Group dynamics in lab courses have been the subject of study beyond university physics courses. For example, important studies in university science, technology, engineering, and mathematics (STEM) courses documenting similar dynamics to the above have been conducted in engineering [12,21], chemistry [22], and biology classes [23,24]. In particular, Donovan *et al.* [15] studied different methods of group formation in a college biology class. These dynamics may contribute to higher attrition rates of women in STEM, given the role of college in the leaky pipeline [25]. Systematic investigations of inequitable group dynamics in a precollege setting are less common. A study by Greenfield [26] found that girls were just as likely to manipulate equipment and record data as boys from elementary to early secondary school. Meanwhile, a study by Jovanovic *et al.* found boys manipulated equipment more than girls in grades 5–8 [5]. Regardless, a broad study by Burkam and Smerdon [27] emphasizes the importance of equitable equipment use for supporting women’s performance in STEM.

In this work, we present a quantitative empirical examination of inequitable group dynamics and of a possible remediation strategy. The context of our work is two introductory physics lab courses for nonmajors at a major public university. Importantly, our work is at the intersection of aforementioned results [4,8], given our courses’ recent redesign implementing much of the inquiry-based framework advocated for in other studies [28,29] at our institution. The size and diversity of these lab courses make them useful testing grounds. The expository in *The Inequality Machine: How College Divides Us* by Tough [30] suggests the observed student body may be of unique interest given its diversity, including dimensions such as race, socioeconomic status, and major [31] explicitly mentioned as limitations or factors of interest in Quinn *et al.* [4].

In this study, we do not empirically assess any link between inquiry-based course design and inequities but rather treat them as motivation for the study of the scaffolding of course elements to ameliorate them.

We choose to focus this work on the gendered aspect of group dynamics for three reasons: One, a number of previous works have also focused on gender [e.g., [4,32]], so this allows direct comparison between student populations. This is important since there is evidence that the effects of a given curriculum can depend on a student’s demographics and other characteristics [33,34]. Two, gender represents a subdivision of students with large populations in two categories. Three, if gender inequities are observed, they may signal broader inequities and provide impetus for follow-up studies examining other dimensions of identity and associated inequities.

The first goal of this work is to extend previous results from Cornell University [4,32] to the context of inquiry-based labs for nonphysics STEM majors at a large public, research-intensive institution in the southern U.S., The University of Texas at Austin. Additionally, by examining student-reported preferences for group work and actual observed behaviors in a single study, we are also able to unify previous results [4,32] by explicitly controlling for preferences in modeling role division in lab activities. The second goal of this work is to provide an assessment of an intervention rooted in social constructivism that is meant to remediate the anticipated inequitable dynamics which involved students completing partner agreement forms.

This paper is organized as follows: In Sec. II we provide theoretical rationale relevant to the study. In Section III we explain the instructional context. In Sec. IV we explain the motivation and implementation of our partner agreement intervention. The next three sections present the methods (Sec. V), results (Sec. VI), and analysis and discussion (Sec. VII) of the two major datasets collected: the preferences survey and the video observations. We conclude by synthesizing our results and their implications both for instruction and for future research on the dynamics of group work in physics lab courses in Sec. VIII.

II. THEORETICAL CONTEXT

A. Constructivism, instructivism, and equity

Many research-based instructional practices common in the physics education research literature fall under the umbrella of the learning theory known as *constructivism*. Constructivism posits that knowledge is constructed by a learner through the active linking together of previous ideas or pieces of information [35]. It contrasts with “traditional” teaching methods, which are based on *instructivist* or *behaviorist* views of learning [36] wherein knowledge is a collection of facts or skills that need to be transmitted from teacher to students [37]. Modern pedagogical

practices such as active learning [2], inquiry-based learning [28,38], interactive engagement [39], and peer-assisted learning [40,41] frequently incorporate students working in groups and are generally rooted in a more specific type of constructivism called *social constructivism*. Social constructivism emphasizes that knowledge construction occurs through social interaction with people [42]. As reviewed in the introduction, there is empirical evidence that pedagogical practices based on social constructivism can enable inequities absent proactive remediation efforts on the part of the instructor [4,8,10–15,32]. Here we define, and will subsequently focus on, equity as structuring course policies that directly consider students' identities and backgrounds, so that current and past structural injustices are addressed to help students learn the course material (similar to [43,44]). Equality, on the other hand, we define as having course policies that treat all students the same regardless of their identities or backgrounds.

There are several theoretical reasons why we might expect some pedagogies grounded on social constructivism to have the capacity to inadvertently reinforce inequitable outcomes. A few well-studied examples which apply specifically to inquiry-based labs include

- **Stereotype threat:** Inquiry-based labs often require students to engage in open-ended exploration and problem-solving, which can lead to increased performance pressure. Stereotype threat, the concern of confirming negative stereotypes about one's group, can be heightened in such situations [45]. This pressure can disproportionately affect systemically nondominant groups, including women, and impact their performance and confidence in the lab setting.
- **Confidence and self-efficacy:** Research suggests that women, on average, may exhibit lower confidence and self-efficacy in STEM fields compared to men [46–48]. Inquiry-based labs can involve higher levels of autonomy, uncertainty, and risk taking, which may lead to, or be impacted by, decreased confidence among students who are less familiar with this type of learning environment. Lower levels of confidence and self-efficacy can influence participation and engagement [49].
- **Identity-based participation patterns:** Lab classes in general involve working in groups and group problem solving. Since identity-based inequities exist in society, for example regarding gender, race, class, and other categories, they can play a role governing the division of labor in group work. This division of labor can perpetuate traditional (e.g., gender) roles and create inequitable participation opportunities and experiences [50,51].
- **Classroom climate and peer or instructor bias:** Group work involves interaction with peers and inquiry-based labs typically require proactive support from instructors. Biases, even unconscious, may influence

interactions with and between students and, even inadvertently, affect the experiences and performance of systemically nondominant groups [52,53].

In the case of lecture courses this issue has been studied extensively [8,10–15,54] and some specific remediation strategies have been proposed, studied, and found to be effective [12,14]. All lab classes which involve group work, similarly, can produce inequities. Based on these theoretical considerations, we expect that inquiry-based labs would be especially susceptible to these problems, and this was found in the aforementioned study [4], which directly compared traditional and inquiry-based labs.² In particular, from both theoretical and empirical angles, we expect group work to be a locus of inequitable dynamics in lab courses in general and inquiry-based labs in particular.

B. Group work practices

To understand how students work together in groups, and how instructors might structure group work to ensure pedagogically sound and equitable outcomes, it is useful to consider three distinct group work practices: instructivist, collaborative, and cooperative learning. A visual depiction of our framework, explained below, is provided in Fig. 1.

Instructivist approaches involve highly structured group work. In theory, this maximizes instructor control over group dynamics. This may have the benefit of allowing the instructor to preclude inequitable outcomes by carefully structuring groups and how groups operate. On the other hand, by at least some measures, instructivist methods can be pedagogically inferior [2,3] to other methods such as those rooted in social constructivism given they do not incorporate active involvement on the part of the student. Especially if effective group work constitutes a learning goal of a course, not just an instructional tool, we would prefer to employ a different framework for organizing group work in class.

Collaborative learning and cooperative learning, meanwhile, are rooted in social constructivism. In collaborative learning, instructors direct students to work together in groups freely, without assigning roles or structuring group work, provided they achieve a certain goal or outcome by dividing tasks effectively among themselves. Cooperative learning is more structured, and involves scaffolding group work so that students work together more cohesively, preferably because the goal or outcome is not achievable by individuals working independently and merely collating their work at the end. According to Davidson [55],

²It is important to note that any extent to which traditional labs may be more equitable than inquiry-based labs in the absence of appropriate scaffolding of group work (or other course elements) is in an important sense superficial. Traditional labs offer less rich and less authentically scientific activities and therefore preclude inequities via rigidity. In short, there is more room for additional tasks, social dynamics, and psychological forces in an inquiry-based lab, and so more room for inequities.

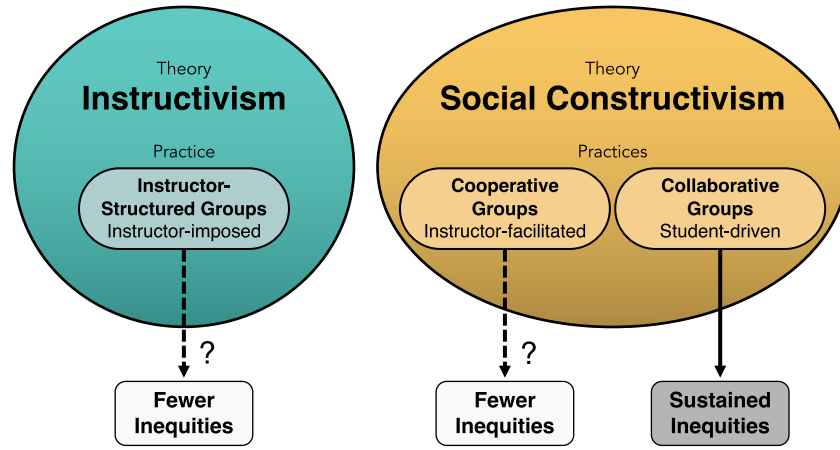


FIG. 1. Pedagogical practices considered in this work depicted within their respective theoretical frameworks. For each pedagogical practice, arrows indicate known link to sustained inequities (solid line) and plausible links to less inequities (dashed lines).

cooperative learning has specific characteristics: a common task or learning activity suitable for group work, small-group interaction, norms for cooperative and mutually helpful behavior among students, individual accountability and responsibility (with a possibility of including group accountability), and positive interdependence. It is important to differentiate the structure provided to students in instructivist and cooperative learning—the former dictates group structure and dynamics while the latter scaffolds students’ self-regulation.

Collaborative learning has the feature of being aligned with social constructivism, but it is highly plausible, and has been argued in the literature [56], that the free-form nature of collaborative group work may enable social and cultural factors to preserve, reproduce, or produce inequities. Group work in inquiry-based labs without explicitly structured group dynamics resides within this framework.

Cooperative learning methods offer both alignment with social constructivism, which is better empirically supported [2,3] and matches the inquiry-based learning framework of our lab courses, as well as specific structures meant to ensure healthy group dynamics. We argue it is therefore an appealing framework for resolving inequities observed in group work in inquiry-based labs.

We will rely on instructivist, collaborative, and cooperative learning, to frame prescriptions for group work as well as our proposed intervention, as explained in the corresponding section (Sec. IV).

III. INSTRUCTIONAL CONTEXT

We investigated two introductory physics lab courses, which took place during the Fall 2022 semester at The University of Texas at Austin. While the two courses are sequential and together constitute a two-semester introductory sequence, we studied students from the two courses

simultaneously; we did not track a cohort of students through both courses. Each course is a single credit hour taken by students from one of three introductory corequisite lecture sequences, including algebra-based physics, calculus-based physics for life science majors, and calculus-based physics for engineering majors. In some cases, students with prior credit are not enrolled in a corequisite lecture course. This setup mixes students from all tracks into the same lab sections and provides an important dimension of diversity in these lab courses.

The first course, which we will refer to here as Physics I Lab, covers standard topics in mechanics. The second, which we will refer to as Physics II Lab, continues with optics, electromagnetism, and some modern physics. Both courses are designed to implement the structured quantitative inquiry framework [57–59], which has some similarities to the investigative science learning environment [60] and scientific community [61] approaches. The structured quantitative inquiry framework provides students with genuine investigative freedom, supported by research-based scaffolding (i.e., invention activities [62]), and requires students to make fully quantitative comparisons of models with data.

Both courses are very large (~1000 students per course) and diverse along several dimensions. At the university level, approximately 20% of students are first-generation college students [63] and a similar percentage are Pell-grant eligible students [64,65], a federal financial aid program open to students with significant financial need. This institution is also designated a Hispanic-serving institution [66]. The students in both lab courses are a representative cross section of this student body, as shown in the demographic breakdown from an anonymous survey in Table I.

Lab sections are taught by graduate teaching assistants (TAs), occasionally with assistance from undergraduate learning assistants (LAs). Each course is supervised by a

TABLE I. Self-reported, anonymous results from a survey of student demographic information: gender (including nonbinary or other options), racial or ethnic identity [67], and parents' highest level of education [68] across courses ($\approx 70\%$ response rate). Racial or ethnic groups were not considered mutually exclusive. Counts may not equal the total as students may not have answered all background questions or preferred to not disclose.

Student-level variables	Survey responses		
	Full sample	Physics I Lab	Physics II Lab
All	1316	788	528
Intervention			
Control	501	325	176
Partner agreements	427	243	184
Gender			
Women	751	485	266
Men	479	255	224
Nonbinary or other	32	18	14
Race or ethnicity			
American Indian or Alaska Native	11	8	3
Asian	462	262	200
Black or African American	87	61	26
Hispanic, Latino, or Spanish	325	197	128
Middle Eastern or North African	42	26	16
Native Hawaiian or Other	3	1	2
Pacific Islander			
White	452	273	179
Some other race or ethnicity	6	4	2
Parents' highest level of education			
High school	129	73	56
Some college but no degree	82	61	21
Associate's or technical degree	50	33	17
Bachelor's degree	359	211	148
Master's degree or above	597	353	244

faculty instructor of record and two to three graduate assistant instructors, or “head TAs.” Head TAs collectively are responsible for helping with curriculum development, running weekly instructional meetings, resolving grade disputes, and supporting the other TAs.

Each course includes nine lab sessions. Each session is three hours long and meets once a week. During a given lab session, students usually work in groups of two, but occasionally three. In rare instances, students work in a larger group or on their own, but that practice is discouraged. Lab activities typically involve designing an experiment to test how well a given model describes a physical system. Student groups collectively turn in a single set of informal, but structured, “lab notes” at the start of the following lab session which document their procedure, analysis, results, and conclusions. This gives students a week to work on analysis and writing outside of class. Because of this, most students spend class time collecting and analyzing data and leave the write-up for outside of class.

Students are allowed to pick their own partners throughout the semester. They change partners or groups every three labs, so that they work with three distinct partners or groups in a given semester. This is occasionally complicated by absences or students dropping the class, in which case groups may be slightly shuffled.

In addition to lab assignments, students individually complete prelab activities and a final capstone quiz or project. The prelab activities are completed before lab sessions as quizzes on the Canvas learning management system and are graded on completion. In Physics I Lab, the final assignment is a lab practical quiz that is meant to test student's mastery of essential measurement methods, analysis tools, and familiarity with equipment. In Physics II Lab, the final assignment involves students proposing and executing their own experiment on a topic of their choice and turning in a scientific poster. The inclusion of these end-of-semester individual assignments, which are worth 20% of students' final grade, are meant to motivate individual responsibility, an important best practice for group work [69].

IV. INTERVENTION

Given the similar pedagogical framework and cultural context, we anticipated observing comparably inequitable group dynamics similar to previous works [4]. As such, we designed an intervention aimed at remediating expected inequities. Below we explain the motivation, form, and implementation of this intervention.

Some prescriptions for improving the equity of group work exist in the literature. For example, a highly structured approach to designing and managing groups is advocated by Heller *et al.* [16,17]. In this framework, students are assigned specific roles which are regularly rotated, are prompted to write reflections on their experiences with their group, and are given group assignments that avoid placing women in the minority. In a summary of effective group work practices for college courses, Rosser's suggestions include ensuring rotation of instructor-defined roles throughout the semester and to avoid isolating women in groups [70].³ In the language of Sec. II, these prescriptions may be thought of as implementations of instructivist approaches with the corresponding benefits and drawbacks.

We seek alternative solutions for a few reasons. First, we believe that teaching students to work effectively in groups is an important learning goal for lab courses in itself. This is consistent with the recommendation of the American

³Rosser also recommends allowing students to select their own leader rather than having one assigned by the instructor, to allow students to assign their own roles initially, and to ensure tasks are group worthy, all of which would be categorized as cooperative group practices. Rosser's foil, the fictional professor Peter Adams, is more completely instructivist and the recommendations of the summary are largely in the direction of cooperative group practices as we advocate here.

Association of Physics Teachers, which categorizes working in small groups as a component of scientific collaboration [1]. We expect that a highly structured, top-down approach to group management rooted in an instructivist approach does not provide students with a sufficiently active role to learn to resolve problematic group dynamics. In the spirit of the “structured quantitative inquiry” philosophy, which aligns with social constructivism and has demonstrated effectiveness for teaching students experimental physics [71], we seek solutions which enhance and scaffold students’ active role in shaping their group work. This allows students to learn to resolve inequities and establish more effective group dynamics. Second, research has shown that highly structured approaches to group work are often met with resistance from students [72]. Third, we prefer to avoid explicit role assignment and rotation because evidence suggests it is better for student learning for them to share, not split, work, even if the splitting is equitable with respect to gender [73]. Fourth, although avoiding isolating women in groups may be sufficient to prevent inequities in collective problem solving [17], it is unclear if this is also sufficient to prevent inequitable divisions of labor (e.g., in equipment use).

We therefore designed an intervention that was meant to give students an active role in preempting and resolving problematic group dynamics themselves. In the language of Sec. II, we aimed to encourage cooperative group work.⁴ This intervention has three components:

- **Individual reflections:** Students were given a one-time, individual writing assignment, to be completed outside of class, before any lab sections met. The assignment asked students to reflect on their values and experiences with group work and to write about them. This component of the intervention is inspired heavily by the values affirmation intervention [74,75], which has been shown to reduce gender achievement gaps on high-stakes exams by combating stereotype threat. The primary purpose of this exercise was to prime students’ awareness of what was important for them in group work, so that they would be better equipped to recognize when their lab experience did not conform to their values for group work and learning. We expected that by providing students with an opportunity to reflect on their values, we would induce more dialogue between group members in lab and through partner agreement or reflection forms. We also speculated it would reduce stereotype threat or other identity-based

issues which could play a role in group dynamics as explained in Sec. II.

- **Partner agreement forms:** Each time students were put into a new group, they were tasked with collectively filling out a partner agreement form. This form required students to introduce themselves and to have an explicit conversation about how work would be split or shared. It was deliberately worded not to bias students towards any particular way of sharing or splitting work, while giving them an opportunity to express the preferences that were primed by the individual reflection assignment. This component of the intervention was motivated in part by results suggesting reduced inequities due to explicit conversations about equipment usage [76]. It also provides a space for norm setting and the development of positive interdependence as components of cooperative group work as explained in Sec. II.
- **Partner reflection forms:** Each time students returned to the same group, they were tasked with collectively filling out a reflection form. This gave them an opportunity to discuss their experience working together the previous week. This element was borrowed from the aforementioned framework from Heller *et al.* [17], since it fits with our preferred approach. Importantly, our partner reflection forms differed in that they were done collectively, not individually.

The individual reflection assignment, partner agreement form, and partner reflection form can be found in the Supplemental Material [77].

Students were given participation credit for completing the individual reflection assignment. They were not given any points for completing the partner agreement nor reflection forms to minimize differences in grading across all sections. Instead, TAs required students to complete these at the start of class before proceeding with lab activities. It is worth noting that since students were not obliged to invest significant effort on these activities, some students may choose not to make maximal use of them. This was deliberate, since whether or not students proactively make use of course structures to obtain equitable and preferred modes of group dynamics is part of what we are testing in this study.

The control sections were not given the individual reflections, partner agreement forms, or partner reflection forms, but were otherwise treated identically to the intervention sections. It is possible that students in the control section may have learned of some aspects of the intervention going on in the sections it was applied to—indeed, we did not hide the study from the students. However, we do not expect this to have impacted our results very much. Small differences between lab sections are common due to differences in TA style and students do not perceive these differences as out of the ordinary.

⁴Strictly speaking, if students utilize partner agreement forms to divide labor in specific ways, this may still be considered collaborative group work, but we will still judge the intervention as successful if it nevertheless reduces inequities.

Partner agreement forms have been implemented and studied before—typically in courses with a project—and these studies have found group contracts often improve communication [72,78–80]. However, few studies discuss role taking. Students in one study were assigned a role at the start of the class [78]. In another study by Chang and Brickman on group work in an introductory biology class [72], students were instructed to rotate roles explicitly as well as to write and follow group contracts. However, students did not assign or rotate roles explicitly and they often disregarded their group contracts. These implementations differ from ours as we wanted students to share their roles in a context where their work varied week to week.

We assess the success of the intervention from the effects on equitable dynamics as observed in the video recordings. We expect that the intervention will eliminate or mitigate gendered role taking and prevent introducing new inequities as compared to the control sections.

V. METHODS

A. Preferences survey

Prior to the first week of lab, students were asked to indicate their preferences for different lab activities, forms of role distribution, and leadership styles [32]. Here, the text and responses of these questions are reproduced.

“Which of the following experiment tasks do you prefer taking on? (Select all that apply)”

- (a) Setting up the apparatus and collecting data.
- (b) Writing up the lab procedures and conclusions.
- (c) Analyzing data and making graphs.
- (d) Managing the group progress.
- (e) No preference or none of the above.

“Which of the following approaches to group tasks do you prefer?”

- (a) One where each person has a different task.
- (b) One where everyone works on each task together.
- (c) One where everyone takes turns with each task.
- (d) No preference.
- (e) Something else.

“Which of the following approaches to leadership do you prefer?”

- (a) One where one student regularly takes on the leadership role.
- (b) One where no one takes on the leadership role.
- (c) One where the leadership role rotates between students.
- (d) No preference.
- (e) Something else.

All three preferences questions were closed response. Multiple preferences could be selected on the activity preferences question. Only one preference could be selected on the role distribution and leadership preferences questions. These questions appeared in one of the mandatory prelab quizzes that students completed

electronically before each lab (see Sec. III). We had 1871 completed responses.⁵

We examined survey results for differences across gender controlling for course, track, and the interaction of course and track in our model. The three lecture tracks act as a proxy for student majors and therefore, motivations which may in turn influence preferences. Additionally, student preferences in Physics II Lab may differ from student preferences in Physics I Lab because students have gained more familiarity with the course structure and the lab’s style of group work. These changes in preferences between courses can also vary with lecture track, as some students’ preferences may evolve differently to better match their priorities. These various conditions are controlled for in the logistic regression model for role preferences given in Eq. (1),

$$\begin{aligned} \text{logit}(R) = & \beta_0 + \beta_1 \text{Woman} + \beta_2 \text{PhysIIILab} + \beta_3 \text{CalcEngr} \\ & + \beta_4 \text{CalcLifeSci} + \beta_5 \text{NoCoreq} \\ & + \beta_6 (\text{PhysIIILab} * \text{CalcEngr}) \\ & + \beta_7 (\text{PhysIIILab} * \text{CalcLifeSci}) \\ & + \beta_8 (\text{PhysIIILab} * \text{NoCoreq}), \end{aligned} \quad (1)$$

where R is the response variable, which is the binary preference selected by students; β_0 is the intercept (Man, Physics I Lab, Algebra-based track); Woman indicates if a student is a woman; PhysIIILab indicates if a student is enrolled in Physics II Lab; CalcEngr indicates if a student is enrolled in the calculus-based physics track for engineering majors; CalcLifeSci indicates if a student is enrolled in the calculus-based physics track for life science majors; and NoCoreq indicates if a student is not enrolled in a corequisite lecture course.

For the activities preference survey, because students could select as many or as few responses as they chose, we treated each of the five responses as binary logistic regressions as in Eq. (1). We identified gendered differences in the activities preferences using the regression estimates. For the role distribution and leadership preferences questions, because students could only select only one response, we treated each question as a multinomial logistic regression controlling for the same factors as shown in Eq. (1). For an introduction to multinomial logistic regression and its uses, see work by Theobald *et al.* [81]. For these role distribution and leadership preferences questions, we used pairwise comparisons of means to identify gendered differences in specific answer choices.

⁵For the preferences survey, we use university-supplied binary data on gender since the survey in Table I, which allowed students to self-report their own gender, was anonymous.

TABLE II. Coding scheme used for video observations. The *laptop*, *calculator*, and *paper* codes were later collapsed.

Code	Description
Equipment	Student was handling the equipment. This includes handling objects that are not necessarily lab equipment (e.g., a phone) when it was explicitly obvious the materials were being used to conduct the experiment (e.g., timing a pendulum's period).
Desktop	Student was operating a lab desktop computer.
Laptop	Student was using a personal computer. This includes iPad or tablet use, but excludes cell phone use.
Calculator	Student was using a calculator. This includes cell phone use when it was explicitly obvious the phone was being used as a calculator.
Paper	Student was using pen and paper.
Other	All other student activities.

B. Video observations

To evaluate how students divide tasks in the setting of a laboratory course, we conducted observations of recorded sections. Out of 93 lab sections covering both Physics I and Physics II Labs during the Fall 2022 semester, we video recorded four sections. These sections included one control and one intervention section for Physics I Lab and one control and one intervention section for Physics II Lab. All sessions for the semester were recorded for these sections.

All four recorded sections were taught by head TA assistant instructors, as opposed to TAs. These sections were chosen for analysis under the assumption that they would be the most uniform subset of sections, as well as most adherent to the course objectives, minimizing instructor effects compared to using novice TAs. The four sections took place at the same time and weekday, and included two sections of the Physics I Lab and two sections of the Physics II Lab, with a control and partner agreement intervention section for each.

Labs started with a brief lecture from the TAs that ranged from roughly 10 to 30 min. During this period, students listened and took notes and did not start on lab work until the lecture concluded. We did not code student activities during this period. Once the TA finished their lecture, we began coding student activity. Every 5 min, a researcher coded what each student was doing at that time according to our coding scheme described below. When it was unclear what a student was doing at these 5 min increments, we checked the video up to 30 sec before and after to make a determination.

We used a coding scheme similar to Quinn *et al.* [4] (see Table II for a description of each code). The *other* code covered a broad range of activities including students being off-task (e.g., using their phone, leaving the room, and talking to peers) as well as on-task (e.g., thinking and discussing with peers, LAs, or TAs). Students not touching but looking at a computer screen they had been scrolling on or typing at within 30 seconds or looking at a piece of paper and holding a pencil without actually writing were coded as the closest relevant activity, rather than *other*. Additionally, when a student was holding equipment, but not using it to explicitly conduct the experiment, we coded it as

equipment. A student who watched another student do an activity was coded as *other*.

In our analysis, we chose to combine the *laptop*, *calculator*, and *paper* codes as the distinction between these activities was unsubstantial in two important ways. First, students often used their personal computers to do calculations and take notes. Second, all three activities were associated with analysis or report-writing and required technical understanding but not physical engagement with lab materials. They were thus functionally similar to one another, but distinct from conducting the experiment itself. The *desktop* code was not included in the grouping of *laptop*, *calculator*, and *paper* because the desktop computer had mixed uses. Students often used the desktop computers to collect data and could therefore be linked to the equipment code. Nevertheless, desktop computers were also often used to read lab instructions, conduct data analysis, or write lab notes which could align desktop computers more with the *laptop*, *calculator*, and *paper* codes. Because of this conflict, we left *desktop* as an independent code.

1. Coders and interrater reliability

To establish interrater reliability, three researchers coded 23 students in a single 3-h recorded class session. For that purpose, we chose the second lab session in the Physics I Lab course. We chose that lab session because students frequently use a diverse set of materials and methods and it would make any difficulties with using the coding scheme apparent. For example, many students use their phones as timers in this lab session; coding this as *equipment* requires more careful observation than equipment observations, such as using a scale, in later labs.

We obtained a Fleiss's kappa [82] of 0.80 and kappas over 0.75 signify excellent agreement [83]. When we combined the *laptop*, *calculator*, and *paper* codes, our Fleiss's kappa increased to 0.84. After coding, the researchers discussed their disagreements and resolved any disputed coded segments by coming to consensus. There were no trends in which codes caused more disagreements. The researchers then coded separate sections. Two of us (M. D. and A. L.) each coded one section of the Physics I Lab and E. H. coded both sections of the Physics II Lab.

TABLE III. Student and observation demographic data from the four lab sections which were recorded. Student demographic data indicates the number of men and women in each section. An observation describes one student in one lab period, thus, observation demographic data indicates the number of men and women in each session across the semester. For example, we have 8 unique students that are men in the Physics I Lab control section; we have 61 observations of these 8 unique students across the full semester due to absences. Note that one student preferred not to share their gender.

	Student demographics				Observation demographics			
	Physics I Lab		Physics II Lab		Physics I Lab		Physics II Lab	
	Control	Intervention	Control	Intervention	Control	Intervention	Control	Intervention
Men	8	13	10	9	61	101	74	67
Women	15	12	8	10	127	103	65	69
Total	24	25	18	19	197	204	139	136

2. Video observations quantitative analysis

While our data was coded in segments, students spent varying amounts of time in the lab. Since our research questions relate to how students work in groups, we normalized observations to a student's group. We refer to this type of data presentation as a student's "group fraction."

A student's group fraction for a coded activity in one class session is the fraction of codes we have of that activity out of the group's total number of codes for that activity in one class session. In a former study, Day *et al.* referred to this as "normalized participation" [6]. For a given student and class session, the group fraction for an activity (g_{activity}) is the number of codes of that student for that activity (N_{activity}) divided by the sum of the total number of codes of that activity (N_{activity}) over all the students in that group, as given in Eq. (2),

$$g_{\text{activity}} = \frac{N_{\text{activity}}}{\sum_{\text{group}} N_{\text{activity}}}. \quad (2)$$

For example, if in one class session we coded a student using equipment twice and their partner using it eight times, the former had an *equipment* group fraction of 0.2 and the latter 0.8. If we did not observe a group doing an activity for an entire class session, no student in that group was assigned a group fraction for that activity. While this is infrequent for codes such as *equipment*, the appearance of codes such as *desktop* varied by group and lab.

A summary of our student population and the total number of class observations can be found in Table III. Most student genders were obtained in an optional supplemental survey given to students who agreed to participate in video recordings. Students were able to self-identify their gender in this survey. A total of 14 students did not fill out this survey, so we supplemented our data with university enrollment information. None of the students who filled out the survey identified as nonbinary or some other gender. Therefore, our dataset only considers men and women.

To analyze our data, we used hierarchical linear modeling. Hierarchical linear modeling is a form of data analysis that accounts for nested structures of data. We use this form

of analysis because we have repeated observations of students in each group. For an introduction to hierarchical linear modeling, see work by Van Dusen and Nissen [84]. For our hierarchical linear model, we treat our data as a two-level model where level-1 data is a group fraction from one class session and level-2 data is a student in a particular group. This allows us to have repeated observations of a student for each group with which they work. We do not consider students across the whole semester as our level-2 data because we want to know how students act within each individual group. Additionally, we do not use a three-level model that takes groups into account as this would violate the assumption of independence of observations, as an increase in one student's group fraction necessitates a decrease in another student's group fraction.

For level-1 data, our *equipment* group fraction for the i th observation of the j th student, $g_{\text{equipment},ij}$, is modeled as in Eq. (3),

$$g_{\text{equipment},ij} = \beta_{0j} + \beta_{1j} \text{GroupSize} + r_{ij}, \quad (3)$$

where β_{0j} is the intercept term, GroupSize is the number of group members associated with the observation, and r_{ij} is the residual term. The level-2 data, which represents a student for each group they join, is modeled as in Eq. (4),

$$\begin{aligned} \beta_{0j} = & \gamma_{00} + \gamma_{01} \text{Woman} + \gamma_{02} \text{Int} \\ & + \gamma_{03}(\text{Int} * \text{Woman}) + \gamma_{04} \text{EquipPref} \\ & + \gamma_{05}(\text{EquipPref} * \text{Int}) + \gamma_{06} \text{CalcEngr} \\ & + \gamma_{07} \text{CalcLifeSci} + \gamma_{08} \text{NoCoreq} + u_{0j}, \end{aligned} \quad (4)$$

where γ_{00} is mean intercept; Woman indicates if the student is a woman; Int indicates if the student was in a group with the partner agreements intervention; EquipPref indicates if the student preferred equipment on the preferences survey; CalcEngr indicates if a student is enrolled in the calculus-based physics track for engineering majors; CalcLifeSci indicates if a student is enrolled in the calculus-based physics track for life science majors; NoCoreq indicates if a student is not enrolled in a corequisite lecture course; and

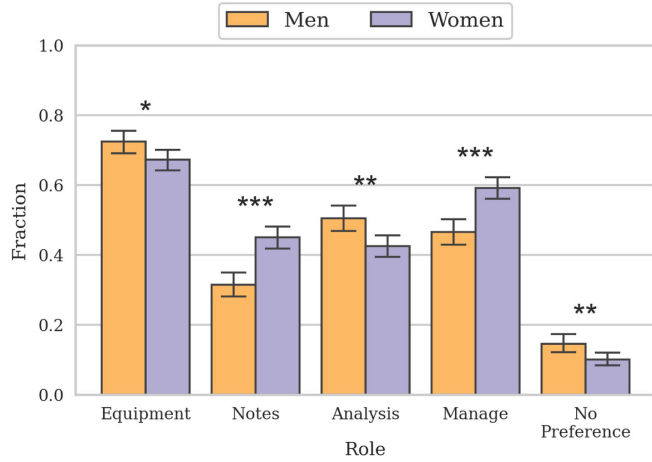


FIG. 2. Expected fraction of men and women that preferred a given role controlling for course, track, and the interaction of course and track. Errors bars represent 95% confidence intervals. The asterisks denote statistical significance where * indicates $p < 0.05$, ** indicates $p < 0.01$, and *** indicates $p < 0.001$. Students could select as many or as few roles as they wanted.

u_{0j} is the residual term. For *laptop*, *calculator*, and *paper* observations, the model is very similar, but the EquipPref term is replaced by two separate terms: one for a preference for notes and one for a preference for analysis (and each with a term interacting with Int). A visual and statistical check ensuring the assumptions of our hierarchical linear model are met can be found in Appendix B.

Additional models accounting for group gender composition were examined as part of this study to see if this had a measurable effect on outcome accounting for preferences. We investigated this as previous research has suggested that women take on different lab roles when in mixed-gender groups [4,7]. Statistical factors which describe the quality of this model, AIC and BIC, indicated that this additional dimension did not improve on Eq. (4). This may be due to lacking a sufficient number of observations from each context. As this more complex model was not a statistical improvement, we do not include its discussion here, but note it may be valuable to examine in future studies.

VI. RESULTS

In this section we briefly present key results from the preferences survey and video observations. We leave further analysis and interpretations to Sec. VII.

A. Preferences survey

The expected fraction of student preferences for certain roles in the lab, controlling for different courses and tracks of physics that students were enrolled in, is shown in Fig. 2. We find that women and men indicate different preferences at the beginning of the semester. Women more often prefer notes ($p < 0.001$) and managing ($p < 0.001$), while men

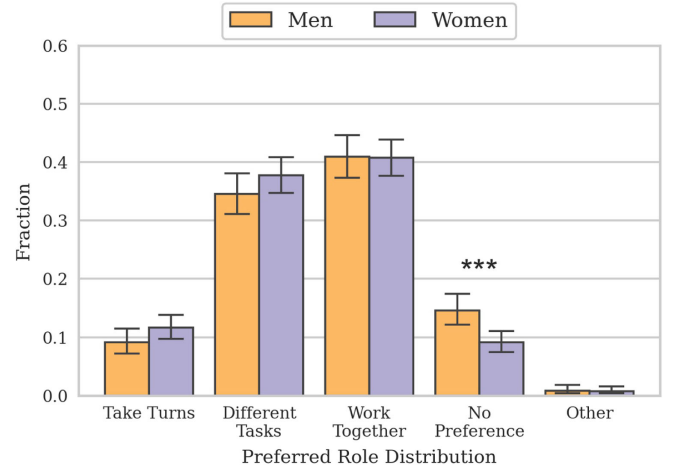


FIG. 3. Expected fraction of men and women that preferred a given method of role distribution controlling for course, track, and the interaction of course and track. Errors bars represent 95% confidence intervals. The asterisks denote statistical significance where * indicates $p < 0.05$, ** indicates $p < 0.01$, and *** indicates $p < 0.001$. Students could select only one answer.

more often prefer equipment ($p = 0.024$), analysis ($p = 0.001$) or have no preferred role ($p = 0.004$). The full numerical results of our regression models are included in Appendix C.

The expected fraction of student preferences for role distributions in lab controlling for course and track is shown in Fig. 3. From a pairwise comparison of means, we find that men are more likely than women to report having no preference ($p = 0.001$).

The expected fraction of student's preferences for role distributions in lab controlling for course and track is shown in Fig. 4. From a pairwise comparison of means, we

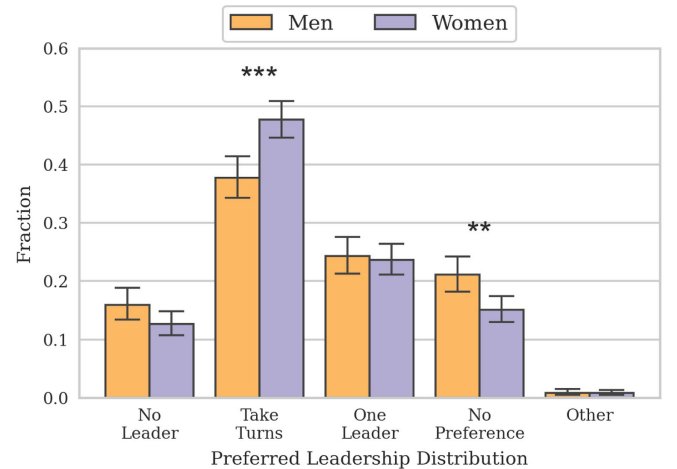


FIG. 4. Expected fraction of men and women that preferred a given leadership style controlling for course, track, and the interaction of course and track. Errors bars represent 95% confidence intervals. The asterisks denote statistical significance where * indicates $p < 0.05$, ** indicates $p < 0.01$, and *** indicates $p < 0.001$. Students could select only one answer.

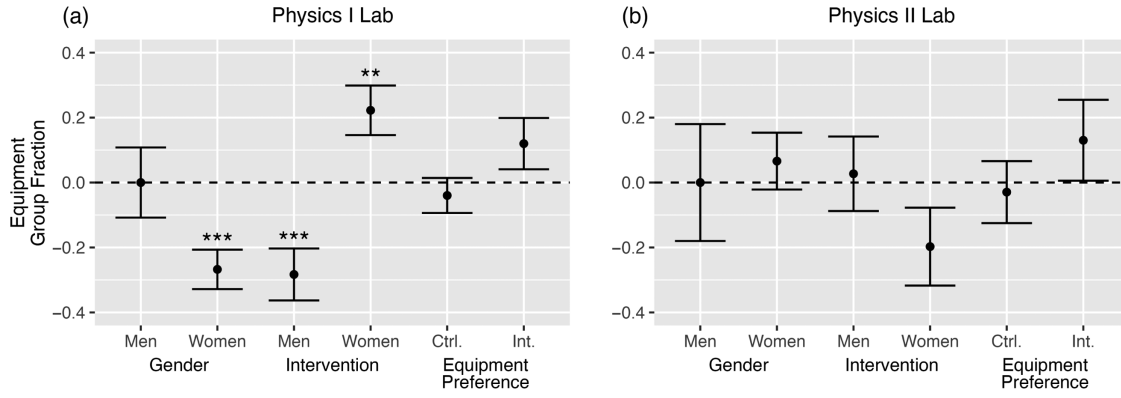


FIG. 5. Results from multilevel regression in (a) Physics I Lab and (b) Physics II Lab for *equipment* group fraction. These results are controlling for group size, lecture track, and random effects. The base term for each course is the group fraction of men in the control section and enrolled in the algebra-based lecture track who did not indicate a preference for equipment. The error bars represent the standard error of the regression coefficients. The asterisks denote statistical significance where * indicates $p < 0.05$, ** indicates $p < 0.01$, and *** indicates $p < 0.001$. The full model output can be found in Table IV.

observe that women more often reported a preference for taking turns in leadership ($p < 0.001$) while men more often reported no preference ($p = 0.004$).

B. Video observations

The results of our regression model for the *equipment* group fraction are shown in Fig. 5 and Table IV. In Physics I Lab without partner agreements, we find that women are responsible for less equipment usage than men ($\beta = -0.226 \pm 0.061$, $p < 0.001$) accounting for group size, lecture track, equipment preference, and random effects. However, in Physics II Lab without partner agreements, we do not observe a gendered difference in equipment usage ($\beta = 0.066 \pm 0.088$, $p = 0.452$).

The partner agreements had differing effectiveness across the two courses. In Physics I Lab, compared to men in the control section, men in the partner agreements section had a lower average *equipment* group fraction ($\beta = -0.283 \pm 0.080$, $p < 0.001$), while women had a higher average *equipment* group fraction ($\beta = 0.222 \pm 0.076$, $p = 0.004$). In Physics II Lab, compared to men in the control section, we did not observe a statistically significant difference in *equipment* group fraction for men or women who used partner agreements.

Notably, indicating a preference for using equipment in Physics I Lab did not lead to a statistically significant increase in *equipment* group fraction for students with or without partner agreements. In Physics II Lab, similarly,

TABLE IV. Results from linear regression for *equipment* group fraction. The table shows the regression coefficient, standard error, and p value (in parentheses). The conditional (marginal) R^2 values for these models are 0.41 (0.16) for Physics I Lab and 0.52 (0.08) for Physics II Lab.

Predictor	Physics I Lab	Physics II Lab
Intercept	1.024 ± 0.108 (<0.001)	0.790 ± 0.180 (<0.001)
Group size	-0.153 ± 0.031 (<0.001)	-0.163 ± 0.068 (0.017)
Track (ref: Algebra-based)		
Calculus-based for Engineers	-0.062 ± 0.059 (0.288)	0.036 ± 0.072 (0.615)
Calculus-based for Life Sciences	0.064 ± 0.047 (0.172)	0.025 ± 0.094 (0.792)
No Corequisite	0.100 ± 0.057 (0.083)	0.060 ± 0.089 (0.496)
Gender (ref: Men)		
Women	-0.267 ± 0.061 (<0.001)	0.066 ± 0.088 (0.452)
Partner Agreements (ref: Control)		
Men	-0.283 ± 0.080 (<0.001)	0.027 ± 0.115 (0.814)
Women	0.222 ± 0.076 (0.004)	-0.197 ± 0.120 (0.101)
Equipment preferred (ref: False)		
Control	-0.040 ± 0.054 (0.461)	-0.029 ± 0.095 (0.758)
Partner agreements	0.120 ± 0.079 (0.130)	0.130 ± 0.125 (0.300)

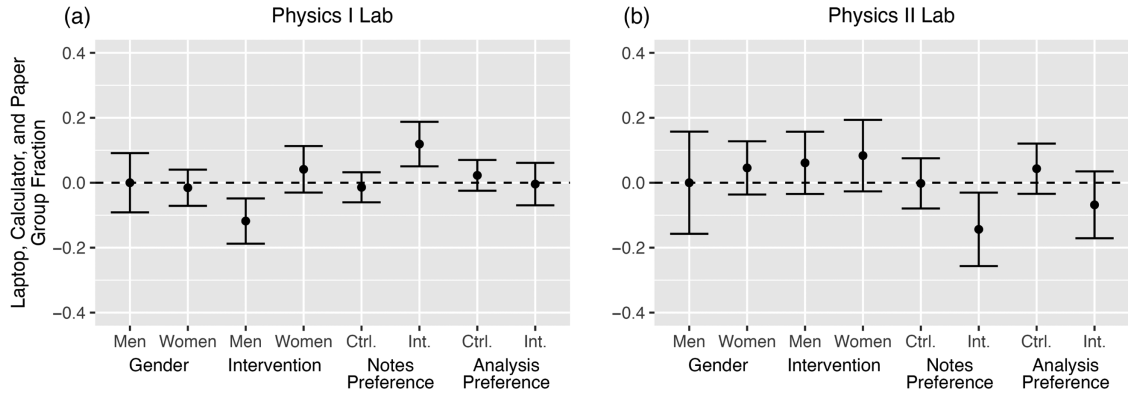


FIG. 6. Results from multilevel regression in (a) Physics I Lab and (b) Physics II Lab for *laptop*, *calculator*, and *paper* group fraction. These results are controlling for group size, lecture track, and random effects. The base term for each course is the group fraction of men in the control section and enrolled in the algebra-based lecture track who did not indicate a preference for notes or analysis. The error bars represent the standard error of the regression coefficients. There were no statistically significant differences. The full model output can be found in Table V.

indicating a preference for equipment did not lead to a statistically significant increase in *equipment* group fraction for students with or without partner agreements.

Figure 6 and Table V show the results of our regression model for *laptop*, *calculator*, and *paper* group fraction. In both Physics I Lab and Physics II Lab without partner agreements, we do not see any statistically significant difference in *laptop*, *calculator*, and *paper* group fraction between men and women. We also see no statistically significant effects from partner agreements on men or women's *laptop*, *calculator*, and *paper* group fraction. Similarly, we see no statistically significant effects from student preferences on *laptop*, *calculator*, and *paper* group

fraction in either course with or without partner agreements.

VII. ANALYSIS AND DISCUSSION

In this section we build on the results presented briefly in the previous section. We analyze and interpret these results and models in the context of our theoretical lens and research questions, presented earlier in the paper.

A. Preferences survey

When surveyed at the beginning of a semester, the most popular lab activity among students was equipment use.

TABLE V. Results from linear regression for *laptop*, *calculator*, and *paper* group fraction. The table shows the regression coefficient, standard error, and p value (in parentheses). The conditional (marginal) R^2 values for these models are 0.55 (0.14) for Physics I Lab and 0.62 (0.10) for Physics II Lab.

Predictor	Physics I Lab	Physics II Lab
Intercept	0.821 ± 0.091 (< 0.001)	0.802 ± 0.157 (< 0.001)
Group size	-0.138 ± 0.027 (< 0.001)	-0.152 ± 0.062 (0.014)
Track (ref: Algebra-based)		
Calculus-based for Engineers	-0.092 ± 0.051 (0.068)	-0.075 ± 0.062 (0.231)
Calculus-based for Life Sciences	0.032 ± 0.043 (0.452)	0.025 ± 0.089 (0.777)
No Corequisite	-0.043 ± 0.051 (0.402)	-0.201 ± 0.091 (0.028)
Gender (ref: Men)		
Women	-0.015 ± 0.056 (0.781)	0.046 ± 0.082 (0.578)
Partner agreements (ref: Control)		
Men	-0.118 ± 0.070 (0.091)	0.061 ± 0.096 (0.524)
Women	0.041 ± 0.072 (0.565)	0.083 ± 0.110 (0.449)
Notes preferred (ref: False)		
Control	-0.014 ± 0.046 (0.761)	-0.002 ± 0.077 (0.978)
Partner agreements	0.119 ± 0.069 (0.083)	-0.144 ± 0.113 (0.205)
Analysis preferred (ref: False)		
Control	0.023 ± 0.047 (0.632)	0.043 ± 0.077 (0.577)
Partner agreements	-0.004 ± 0.065 (0.949)	-0.068 ± 0.103 (0.509)

Men indicated a preference for equipment usage slightly more often than women. The difference in the expected fractions between men and women resembles the magnitude of the difference found by Holmes *et al.* [32], although they did not conduct tests of statistical significance on their dataset. Men were also more likely to prefer the analysis role or to indicate having no role preference. Women more often expressed a preference for note taking and managing at the start of the semester. This may be related to the “Hermione” and “secretary” archetypes from Doucette *et al.* [7], suggesting that previously observed gendered division of labor may be driven in part by student preferences.

For student preferences in role distributions, men and women in our courses have similar preferences, albeit men are more likely to have no preference. Generally, students nearly equally prefer working on different tasks or working together on the same tasks. In the language of Sec. II, this could suggest that students are similarly likely to prefer collaborative or cooperative modes of group work. This is notably different from the findings of Holmes *et al.* [32], where both men and women preferred working on the same task together in a laboratory class targeted towards physics majors. This comparatively strong preference among students in our study for splitting the work may be because these students, not being physics majors as was the case in Holmes *et al.* [32], prioritize efficiently completing lab work over content mastery. Another difference with previous results appears in the leadership preferences. In Holmes *et al.* [32], students were unreceptive to a singular leader and were comparatively more likely to prefer having no leader. A large fraction of our study’s students want some form of leadership, whether that is a rotating or singular leader.

The observations of the last paragraph have important implications, since student preferences inevitably intermix with course structures and interventions to produce outcomes. Differences in student preferences between populations suggest best practices for group work may require some institution-specific or course-specific tailoring.

B. Video observations

1. Control and partner agreements

In the nonintervention sections, we observed men being responsible for more equipment usage in their groups than women in Physics I Lab, but not in Physics II Lab. However, we did not find gendered differences in how men and women used laptops, calculators, and paper in their groups. Students primarily used their laptops for data analysis and note taking, while they near exclusively used calculators for analysis and paper for notes. This suggests that, in terms of roles, men were more likely to be a group’s equipment user. We cannot claim that men or women are more or less likely to be note takers or data analysts.

Men being more likely to be their group’s equipment user echoes the results of previous studies that have examined student roles in physics labs. In observations of similarly structured inquiry-based physics labs, Quinn *et al.* [4] found that men were more often responsible for equipment usage, however, they also found women used laptops more. Another study of the labs at the same university found that equipment usage was similarly gendered for in-person courses, but that online courses with fixed groups across the semester resolved this inequity [76].

We found differences compared with a study by Day *et al.* [6] who analyzed role distribution among mixed-gender pairs of students. While their coding system differed from ours in that they had codes just for equipment, computer, and everything else (also called *other*), they found that men and women used equipment a similar amount. However, they observed men more frequently using computers and women were more frequently coded *other*. Since students submitted notes on paper in that course, this suggests that men in the course did more data analysis while women did more note taking.

Recall that in Sec. IV we provided a consideration for assessing the effectiveness of the intervention. It is effective if it resolves or at least mitigates any gender inequitable division of roles and does not introduce any new gendered inequities. By this metric, the results are positive regarding the effectiveness of the partner agreement intervention.

Our results suggest that the partner agreements led to more equitable equipment usage among students in Physics I Lab. For *laptop*, *calculator*, and *paper* in Physics I Lab, the partner agreements did not alter the gendered distribution of labor. Similarly, when we consider Physics II Lab, we see no statistically shifts for *equipment* group fraction or *laptop*, *calculator*, and *paper* group fraction.

While partner agreements do not explicitly prompt students to consider gender equitable labor, their effectiveness for equipment usage in Physics I Lab has a possible explanation. In Physics I Lab, students are required to complete a practical quiz at the end of the course which tests, among other things, skills with equipment. When completing partner agreements, students may be more motivated to ensure everyone gets equal experience with the equipment or more motivated to self-advocate for a role in equipment use.

The findings of Zhang *et al.* on the effects of formal contracts and competence trust on group work might offer an additional lens on the outcomes observed here [85]. Competence trust is how confident a student is in their partners’ experience and ability to complete a task at a high level [86]. Zhang *et al.*’s findings show maximal benefit to group work when group contracts are present and groups have mild competence trust, where neither partner is perceived to be notably more or less capable by the other.

In the courses studied here, competence trust may be activity-specific; students may inherently perceive some others as more or less competent in Physics I Lab due to their experiences with and perceptions of the subject, equipment, or processes. In Physics II Lab, however, students may perceive each other as being more uniformly capable for two reasons. They may perceive everyone as *less* capable due to the sophisticated, and unfamiliar, electronics in Physics II Lab. Alternatively, they may perceive everyone as *more* capable because they all completed Physics I Lab and have gained familiarity with group work in a university physics lab context. Similarly, students in Physics II Lab may have more college experience or higher maturity levels which could impact students' perceptions of each other's capabilities and influence the expression of their own interests.

Despite students not being given a grade incentive, the changes in equipment usage from the intervention suggest students are engaging with and using the partner agreements. It is possible that students use the conversations prompted by the partner agreements as a tool to complete group work more efficiently, but the results noted here show a more equitable outcome—at least for equipment usage—as a result of their inclusion in the course.

2. Role preferences

Across the Physics I and Physics II Labs, we found that student preferences at the start of the semester did not correlate with their frequency of engaging in observed lab activities. So students who identified as preferring to use equipment more were not later observed to use equipment more often. Interestingly, this behavior was consistent across both control and partner agreement sections.

This common trend of student preferences not affecting actual roles suggests students are dividing roles for other reasons. Previous research has found that students often informally take on lab roles [4,32]. This suggests that informal role assignments are not even due to students with proclivities towards certain work instinctively taking it up. In their study of a project-based physics lab, Stump *et al.* [19] found a managerial role allowed some women a form of self-expression which could promote identity development, engagement, and learning. If students are not taking on roles they want to do, it could potentially impact their learning outcomes and affect (e.g., attitudes).

C. Limitations

While student preferences for roles were probed in the survey, our observation protocol did not directly observe students within all of these roles. Notably, we did not directly observe note taking and data analysis; we observed students using laptops and calculators and writing on paper. Because these activities encompass both note taking and

data analysis, we cannot fully glean how students divided roles. It is possible that some students tended to take on more secretarial or lead scientist roles, however, we could not observe these differences.

It is also important to note that neither the manipulation of equipment nor note taking should be equated with the entirety of scientific practice. Both are necessary components, and they are aspects which are learning goals of the course which have corresponding graded assignments (like the lab notes for each lab or the end of the semester lab practical and final projects). But other activities beyond the scope of our video coding scheme such as problem posing, discussions of experimental design, and data analysis, are also important components. We do not, therefore, have a complete portrait of the inequities in this lab course, even though the differences in equipment usage which are not reducible to preferences implies inequities for at least this aspect according to the definition adopted in Sec. II.

Since the interventions were applied to separate sections and implemented for an entire semester, the data compare sections with potential for different random effects. Hierarchical linear modeling accounts for random differences in behavior of students across their sessions with one group, however, there are possible random differences in sections. We could not account for these because we only have one section per condition (i.e., course and control or intervention). We attempted to minimize the differences between sections by analyzing sections taught by head TAs occurring on the same weekdays at the same times; uncontrollable factors can cause otherwise identical sections to differ in significant ways which could have altered the apparent intervention effectiveness. TA behavior and identity also differed between sections. This could have had an impact on the intervention's effectiveness. However, previous research finds little impact due to instructor gender, suggesting this is unlikely [87–90], although this remains an open question.

In this analysis we have neglected to discuss a “group manager” role. This is because observations were conducted on video that lacked audio. This put identifying a group manager outside the possible scope of this study. Other studies have analyzed the positive [19] and negative [7] aspects group management can have on a woman's experience in physics labs.

We have not controlled for the possibility that students may have chosen to work with students with whom they were previously acquainted with and acknowledge that this could play a role in outcomes. Group formation procedures were the same between control and intervention sections, so this does not impact the assessment of our intervention in its context, but this is an important factor worth considering in other contexts or future studies. See Pulgar *et al.* [91] for one example of a study investigating this dynamic.

We have only considered groups of sizes two and three. Other lab courses may have larger groups. Since this affects the available number of roles per student, it could conceivably produce different outcomes.

VIII. CONCLUSIONS

In this section, we summarize our findings holistically in light of our research goals, draw some conclusions, indicate implications for instruction, and suggest avenues for future work.

The first goal of this study was to apply the same methods as previous work [4,32] in the context of our courses. Given the similarity in course framework, we expected we would observe similar inequities. We did, indeed, observe inequities [4] and they do not appear to be reducible to differences in lecture tracks or role preferences among our students [32]. Given the differences in student populations between those studies and ours, we take this as evidence that they are common to inquiry-based labs or lab courses more broadly. Importantly, by including both preferences and observations in a single study, and by including preferences as a component of our model, we unify and therefore strengthen the conclusion of these studies that preferences do not account for differences in observed lab activities. In fact, we found role preferences played very little part at all in determining actual role taking in labs, which further suggests students may require scaffolded group work to ensure their participation reflects their interests and/or is more equitable.

This emphasizes the tension between best practice pedagogical methods and efforts to promote diversity, equity, and inclusion. An important point of Quinn *et al.* [4] is that these inequities are not merely background effects of culture on all physics courses, they are consequences which can be inadvertently reinforced by particular choices of curriculum design.

Although we did see some relatively small presemester gendered differences in preferences that match Holmes *et al.* [32], we did find some differences in preferences between the two student populations. Our students are more likely to prefer dividing tasks and a single leader. Our observations, however, indicate that these student preferences likely do not result in statistically significant differences in observed behavior. In the language of Sec. II, this suggests that students are willing to engage in cooperative group structures with a brief, recurring intervention that does not explicitly compel them to equitably divide group tasks.

In this study we have only examined inequities based on student gender. There may be inequities based on other demographic criteria such as race or ethnicity or students' academic backgrounds [8]. We plan to follow up with future work to explore these possibilities. Given the differences between Physics I Lab and Physics II Lab, it

would also be interesting to conduct a future study in which the preferences survey is administered both presemester and postsemester to observe how lab activities may influence preferences. Additionally, a future study could examine preferences and observations within each group to explore if preferences have different effects at the start of the semester or the first sessions of new lab groups.

The second goal of this study was to implement and evaluate the impact of an intervention meant to reduce inequitable task division. This intervention was intended to work by scaffolding group work to enhance student's active role in shaping group dynamics, ideally producing something closer to cooperative rather than collaborative group work. Our results show that the intervention had positive results when students had motivation (e.g., a summative practical quiz) to self-advocate for an equal role in equipment usage, since observations showed improvements where inequitable task division existed.

While our results are promising that small interventions can help mitigate inequitable group dynamics, further study is needed to investigate factors noted above, and to examine if these outcomes are true in other contexts. In particular, developing interventions to further instantiate students' self-interest in exploring all roles in a lab course would be beneficial. Using these interventions to broaden student awareness of best practices may improve their intrinsic motivation and promote equitable and effective learning experiences.

A motivation for this work were the theoretical reasons and empirical evidence that inquiry-based labs may enable or inadvertently reinforce inequities. But we have not tested this relationship in this work and, importantly, our conclusions may apply more broadly to traditional labs as well. In fact, our results emphasize the importance of the distinction between inquiry-based labs as a general laboratory design strategy and the structure and scaffolding of group dynamics (among other aspects of the course). Although there may be reasons to believe that inquiry-based design have the potential to enable inequities, our results show that this can be affected by how group dynamics are managed by instructors. Because inequities exist in society, instructors need to intentionally design lab experiences that scaffold and direct groups so as to create equitable experiences regardless of other curriculum choices. Indeed, the richer and more authentically scientific lab activities are, the more this is likely to be both necessary and valuable.

ACKNOWLEDGMENTS

We would like to thank Natasha Holmes, Kathryn Hendren, Jane Huk, Vernita Gordon, and John Yeazell for many useful discussions and feedback. We would like to thank Ben Costello (and his supply-room workers) for

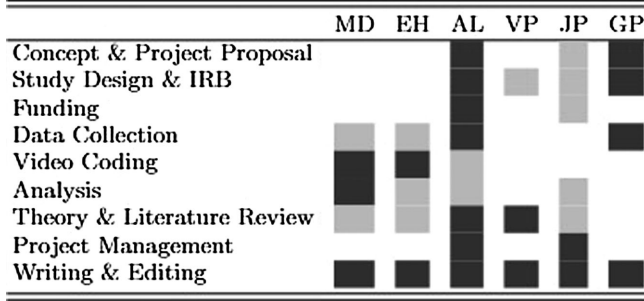


FIG. 7. Contributions of authors. Dark shading corresponds to a major contribution, while a lighter shading corresponds to a minor contribution.

implementing the video recording hardware and managing the software and servers. We would also like to thank the College of Natural Sciences at The University of Texas at Austin for providing funding for this project through the 21st Century Curriculum Redesign Effort, and Kristin Patterson and Keely Finkelstein for their assistance with applying for and administering these funds. We would also like to thank the many students who agreed to participate in this study and the graduate student teaching assistants and undergraduate learning assistants who assisted with teaching the courses.

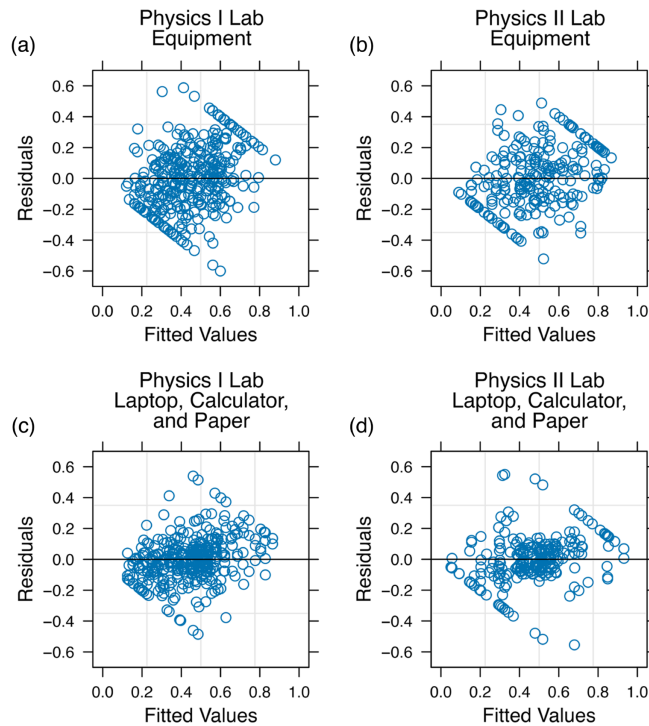


FIG. 8. Visual check for the assumption of linearity for hierarchical linear modeling. These plots show the residuals vs the fitted values for (a) *equipment* group fraction in Physics I Lab, (b) *equipment* group fraction in Physics II Lab, (c) *laptop, calculator, and paper* group fraction in Physics I Lab, and (d) *laptop, calculator, and paper* group fraction in Physics II Lab.

APPENDIX A: AUTHORSHIP MATRIX

Because of the complexity of this project, frequently used author ordering conventions seemed inadequate to properly recognize contributions from all authors. The author ordering of this paper is alphabetical, with an anchor author for the principal investigator. The matrix shown in Fig. 7 describes each author's primary or secondary contribution to significant parts of this project from inception to publication.

APPENDIX B: ASSUMPTIONS FOR HIERARCHICAL LINEAR MODELING

In this Appendix, we discuss how well assumptions were met for hierarchical linear modeling (HLM). Specifically, following guidance from Van Dusen and Nissen, we examine linearity, homogeneity of variance, and normality [84]. Examining linearity, Fig. 8, we observed no trend in these data about the x axes. While there are ceiling and floor effects (shown by the diagonal lines at the top right and bottom left of each plot), there is no overall structure or trend to these data as a whole. To test homogeneity of variance, ANOVA was used across all groups and students. No statistically significant differences were found ($p \sim 1$ in all four cases) indicating that variance was homogeneous.

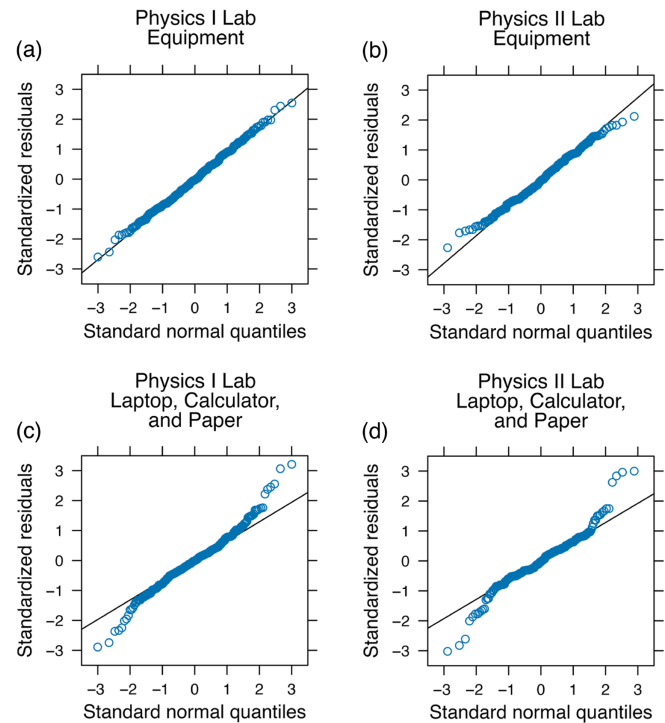


FIG. 9. Visual check for the assumption of normality for hierarchical linear modeling. These plots show the residuals vs the fitted values for (a) *equipment* group fraction in Physics I Lab, (b) *equipment* group fraction in Physics II Lab, (c) *laptop, calculator, and paper* group fraction in Physics I Lab, and (d) *laptop, calculator, and paper* group fraction in Physics II Lab.

Residuals for our models are shown in Fig. 9. Visual inspection shows that our data generally meets the assumption of normality, falling on or near the line in all cases. While there are some deviations at either end of the plot, particularly for the third and fourth models, this does not generally impact p values or estimates [71,92].

For our hierarchical linear modeling, we used the `lmer` function from the *lme4* package in R [93].

APPENDIX C: PREFERENCE SURVEY RESULTS

We present the full results of the preferences survey from Sec. VIA as well as the number of students for each

category, Table VI. Table VII provides the roles students preferred, Table VIII provides the role distributions students preferred, and Table IX provides the leadership distributions students preferred.

To conduct logistic regression for the roles students preferred, we used the `glm` function in the *base* package in R [94]. For role distribution and leadership distribution questions, we conducted multinomial logistic regression using the `multinom` function in the *nnet* package [95]. For all of these questions, we then used the *effects* package to display the probabilities of men and women selecting each answer [96]. For pairwise comparisons of means, we used the `emmeans` function from the *emmeans* package [97].

TABLE VI. Number of men and women enrolled in each course and track for Physics I Lab and Physics II Lab.

Track	Physics I Lab		Physics II Lab	
	Men	Women	Men	Women
Algebra-based	135	332	54	134
Calculus-based for Engineers	119	56	182	87
Calculus-based for Life Sciences	101	218	90	133
Not enrolled in Corequisite	46	71	65	48

TABLE VII. Results from the logistic regressions for student role preferences given in Eq. (1) which controls for gender, course, track, and the interaction of course and track. The table shows the regression coefficient, standard error, p value (in parentheses), and odds ratio (in brackets).

Predictor	Equipment	Notes	Analysis	Managing	No Preference
Intercept	1.033 ± 0.129 (<0.001) [2.809]	-1.025 ± 0.125 (<0.001) [0.359]	-0.199 ± 0.118 (0.091) [0.819]	0.179 ± 0.119 (0.133) [1.196]	-1.946 ± 0.184 (<0.001) [0.143]
Gender (ref: Man)					
Woman	-0.245 ± 0.109 (0.024) [0.782]	0.579 ± 0.105 (<0.001) [1.785]	-0.322 ± 0.100 (0.001) [0.724]	0.510 ± 0.100 (<0.001) [1.666]	-0.429 ± 0.151 (0.004) [0.651]
Course (ref: Physics I Lab)					
Physics II Lab	-0.357 ± 0.182 (0.049) [0.700]	0.676 ± 0.177 (<0.001) [1.965]	0.040 ± 0.177 (0.823) [1.040]	-0.349 ± 0.176 (0.048) [0.706]	0.408 ± 0.264 (0.121) [1.504]
Track (ref: Algebra-based)					
Calculus-based for Engineers	0.022 ± 0.202 (0.915) [1.022]	-0.008 ± 0.197 (0.968) [0.992]	0.592 ± 0.184 (0.001) [1.807]	-0.446 ± 0.184 (0.016) [0.640]	0.076 ± 0.287 (0.791) [1.079]
Calculus-based for Life Sciences	0.026 ± 0.160 (0.873) [1.026]	0.017 ± 0.153 (0.911) [1.017]	0.134 ± 0.148 (0.364) [1.144]	-0.162 ± 0.150 (0.280) [0.850]	-0.006 ± 0.246 (0.980) [0.994]
No Coreq	0.096 ± 0.231 (0.677) [1.101]	0.266 ± 0.214 (0.214) [1.305]	0.275 ± 0.209 (0.189) [1.316]	-0.262 ± 0.211 (0.214) [0.769]	0.020 ± 0.344 (0.954) [1.020]
Course * Track (ref: Physics I Lab * Algebra-based)					
Physics II Lab * Calculus-based for Engineers	0.317 ± 0.283 (0.264) [1.372]	-0.580 ± 0.276 (0.035) [0.560]	-0.215 ± 0.264 (0.414) [0.806]	0.247 ± 0.264 (0.349) [1.280]	-0.158 ± 0.392 (0.686) [0.853]
Physics II Lab * Calculus-based for Life Sciences	0.209 ± 0.263 (0.427) [1.232]	0.014 ± 0.252 (0.957) [1.014]	0.173 ± 0.249 (0.488) [1.189]	-0.257 ± 0.251 (0.305) [0.773]	0.069 ± 0.375 (0.854) [1.071]
Physics II Lab * No Coreq	0.014 ± 0.341 (0.967) [1.014]	-0.435 ± 0.323 (0.178) [0.647]	-0.210 ± 0.320 (0.510) [0.810]	0.414 ± 0.321 (0.197) [1.513]	-0.118 ± 0.487 (0.809) [0.889]

TABLE VIII. Results from the multinomial logistic regression for student role distribution preferences similar to Eq. (1) and detailed in Sec. VA which controls for gender, course, track, and the interaction of course and track. The table shows the regression coefficient, standard error, and p value (in parentheses).

Predictor	Different tasks vs take turns	Work together						No preference vs different tasks	Other vs different tasks	No preference vs work together	Other vs work together	Other vs no preference
		Work together vs take turns	No preference vs take turns	Other vs take turns	Work together vs different tasks	No preference vs different tasks	Other vs different tasks					
Intercept	1.475±0.222 (<0.001)	1.913±0.216 (<0.001)	0.538±0.266 (0.043)	-3.461±1.077 (0.001)	0.438±0.133 (0.001)	-0.937±0.204 (<0.001)	-4.937±1.064 (<0.001)	-1.375±0.197 (<0.001)	-5.374±1.063 (<0.001)	-3.998±1.074 (<0.001)		
Gender (ref: Man)												
Woman	-0.153±0.177 (0.387)	-0.247±0.175 (0.158)	-0.713±0.212 (0.001)	-0.313±0.524 (0.551)	-0.094±0.114 (0.411)	-0.560±0.166 (0.001)	-0.160±0.507 (0.753)	-0.466±0.164 (0.004)	-0.066±0.507 (0.896)	0.400±0.521 (0.442)		
Course (ref: Physics I Lab)												
Physics II Lab	-0.211±0.290 (0.468)	-0.808±0.292 (0.006)	-0.324±0.377 (0.390)	1.162±1.251 (0.353)	-0.597±0.198 (0.003)	-0.114±0.310 (0.713)	1.374±1.233 (0.265)	0.483±0.312 (0.121)	1.971±1.233 (0.110)	1.488±1.255 (0.236)		
Track (ref: Algebra-based)												
Calculus-based for Engineers	-0.034±0.354 (0.924)	-0.060±0.343 (0.862)	-0.020±0.425 (0.962)	2.314±1.178 (0.050)	-0.026±0.207 (0.900)	0.014±0.324 (0.967)	2.349±1.146 (0.040)	0.039±0.312 (0.899)	2.374±1.143 (0.038)	2.334±1.170 (0.046)		
Calculus-based for Life Sciences	-0.408±0.252 (0.106)	-0.663±0.246 (0.007)	-0.305±0.320 (0.339)	1.305±1.139 (0.252)	-0.256±0.166 (0.125)	0.102±0.263 (0.696)	1.714±1.125 (0.128)	0.358±0.257 (0.163)	1.969±1.124 (0.080)	1.610±1.142 (0.158)		
No Corequisite	-0.894±0.330 (0.007)	-1.016±0.317 (0.001)	-0.336±0.396 (0.397)	0.613±1.441 (0.670)	-0.122±0.251 (0.626)	0.558±0.346 (0.106)	1.507±1.428 (0.291)	0.681±0.333 (0.041)	1.630±1.425 (0.253)	0.949±1.444 (0.511)		
Course * Track (ref: Physics I Lab * Algebra-based)												
Physics II Lab *	0.118±0.468 (0.800)	0.187±0.465 (0.687)	0.584±0.569 (0.305)	-2.546±1.556 (0.102)	0.069±0.301 (0.819)	0.465±0.444 (0.295)	-2.666±1.516 (0.079)	0.397±0.441 (0.368)	-2.734±1.515 (0.071)	-3.130±1.549 (0.043)		
Engineers												
Physics II Lab *	0.818±0.424 (0.054)	1.222±0.426 (0.004)	0.771±0.537 (0.151)	-1.757±1.701 (0.301)	0.404±0.281 (0.150)	-0.047±0.431 (0.913)	-2.578±1.671 (0.123)	-0.451±0.432 (0.297)	-2.981±1.671 (0.074)	-2.531±1.703 (0.137)		
Life Sciences												
Physics II Lab *	1.335±0.558 (0.017)	1.794±0.551 (0.001)	0.967±0.670 (0.149)	0.844±1.755 (0.631)	0.459±0.371 (0.216)	-0.368±0.531 (0.488)	-0.491±1.707 (0.774)	-0.827±0.523 (0.114)	-0.950±1.705 (0.577)	-0.124±1.746 (0.943)		
No Corequisite												

TABLE IX. Results from the multinomial logistic regression for student leadership preferences similar to Eq. (1) and detailed in Sec. VA which controls for gender, course, track, and the interaction of course and track. The table shows the regression coefficient, standard error, and p value (in parentheses).

Predictor	Take turns vs no leader	One leader vs no leader	No preference vs no leader	Other vs no leader	One leader vs take turns	No preference vs take turns	Other vs take turns	No preference vs one leader	Other vs one leader	Other vs no preference
Intercept	0.962 ± 0.178 (<0.001)	0.432 ± 0.195 (0.027)	0.098 ± 0.212 (0.643)	-1.677 ± 0.405 (<0.001)	-0.530 ± 0.149 (<0.001)	-0.864 ± 0.171 (<0.001)	-2.640 ± 0.385 (<0.001)	-0.334 ± 0.189 (0.077)	-2.110 ± 0.394 (<0.001)	-1.776 ± 0.402 (<0.001)
Gender (ref: Man)										
Woman	0.467 ± 0.153 (0.002)	0.205 ± 0.165 (0.215)	-0.102 ± 0.175 (0.559)	0.194 ± 0.384 (0.613)	-0.262 ± 0.127 (0.040)	-0.569 ± 0.141 (<0.001)	-0.272 ± 0.369 (0.460)	-0.307 ± 0.154 (0.046)	-0.010 ± 0.375 (0.978)	0.297 ± 0.379 (0.434)
Course (ref: Physics I Lab)										
Physics II Lab	-0.259 ± 0.264 (0.326)	-0.242 ± 0.295 (0.412)	0.192 ± 0.311 (0.537)	0.157 ± 0.522 (0.763)	0.017 ± 0.228 (0.941)	0.451 ± 0.249 (0.071)	0.416 ± 0.487 (0.392)	0.434 ± 0.282 (0.124)	0.399 ± 0.505 (0.429)	-0.035 ± 0.514 (0.946)
Track (ref: Algebra-based)										
Calculus-based for Engineers	0.027 ± 0.275 (0.923)	-0.223 ± 0.313 (0.476)	0.037 ± 0.331 (0.911)	-0.027 ± 0.595 (0.964)	-0.250 ± 0.241 (0.301)	0.011 ± 0.266 (0.968)	-0.053 ± 0.561 (0.924)	0.260 ± 0.305 (0.393)	0.196 ± 0.581 (0.736)	-0.064 ± 0.591 (0.913)
Calculus-based for Life Sciences	0.344 ± 0.245 (0.160)	0.358 ± 0.266 (0.179)	0.422 ± 0.292 (0.147)	-0.496 ± 0.613 (0.418)	0.014 ± 0.182 (0.940)	0.078 ± 0.218 (0.720)	-0.840 ± 0.581 (0.149)	0.065 ± 0.242 (0.790)	-0.854 ± 0.591 (0.149)	-0.918 ± 0.603 (0.128)
No Corequisite	0.198 ± 0.342 (0.562)	0.181 ± 0.374 (0.628)	0.482 ± 0.394 (0.220)	-11.810 ± 0.433 (<0.001)	-0.017 ± 0.263 (0.949)	0.284 ± 0.292 (0.330)	-12.450 ± 0.416 (<0.001)	0.301 ± 0.329 (0.360)	-11.823 ± 0.423 (<0.001)	-12.099 ± 0.424 (<0.001)
Course * Track (ref: Physics I Lab * Algebra-based)										
Physics II Lab *	-0.256 ± 0.391 (0.512)	0.334 ± 0.436 (0.445)	-0.110 ± 0.456 (0.809)	-0.996 ± 0.884 (0.260)	0.590 ± 0.346 (0.088)	0.146 ± 0.371 (0.695)	-0.741 ± 0.843 (0.380)	-0.444 ± 0.419 (0.289)	-1.330 ± 0.865 (0.124)	-0.886 ± 0.875 (0.312)
Calculus-based for Engineers										
Physics II Lab *	-0.636 ± 0.385 (0.098)	-0.113 ± 0.417 (0.786)	-0.557 ± 0.447 (0.212)	-1.732 ± 1.257 (0.168)	0.523 ± 0.316 (0.098)	0.079 ± 0.356 (0.825)	-1.097 ± 1.228 (0.372)	-0.444 ± 0.390 (0.255)	-1.619 ± 1.238 (0.191)	-1.174 ± 1.249 (0.347)
Calculus-based for Life Sciences										
Physics II Lab *	-0.301 ± 0.501 (0.549)	0.008 ± 0.546 (0.988)	-0.201 ± 0.560 (0.720)	11.111 ± 0.433 (<0.001)	0.309 ± 0.414 (0.456)	0.100 ± 0.434 (0.818)	11.854 ± 0.416 (<0.001)	-0.209 ± 0.484 (0.666)	10.935 ± 0.423 (<0.001)	11.119 ± 0.424 (<0.001)
No Corequisite										

- [1] D. MacIsaac, Report: AAPT recommendations for the undergraduate physics laboratory curriculum, *Phys. Teach.* **53**, 253 (2015).
- [2] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth, Active learning increases student performance in science, engineering, and mathematics, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8410 (2014).
- [3] A. I. Ibrahim, N. Sulaiman, and I. Ali, Simultaneous multidimensional impacts of active learning revealed in a first implementation in the MENA region, *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2108666119 (2022).
- [4] K. N. Quinn, M. M. Kelley, K. L. McGill, E. M. Smith, Z. Whipps, and N. G. Holmes, Group roles in unstructured labs show inequitable gender divide, *Phys. Rev. Phys. Educ. Res.* **16**, 010129 (2020).
- [5] J. Jovanovic and S. S. King, Boys and girls in the performance-based science classroom: Who's doing the performing?, *Am. Educ. Res. J.* **35**, 477 (1998).
- [6] J. Day, J. B. Stang, N. G. Holmes, D. Kumar, and D. A. Bonn, Gender gaps and gendered action in a first-year physics laboratory, *Phys. Rev. Phys. Educ. Res.* **12**, 020104 (2016).
- [7] 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).
- [8] V. Gordon, J. Helwig, and J. Huk, *APS March Meeting 2021* 66 (2021).
- [9] D. Jenkins, Women of Color's Experiences and Strategies in Constructing Nonexecutive Community College Leadership: A Case Study, Ph.D. thesis, University of Phoenix, 2017.
- [10] S. M. Aguilon, G.-F. Siegmund, R. H. Petipas, A. G. Drake, S. Cotner, and C. J. Ballen, Impact of cold-calling on student voluntary participation, *CBE Life Sci. Educ.* **19**, ar12 (2020).
- [11] E. J. Dallimore, J. H. Hertenstein, and M. B. Platt, Impact of cold-calling on student voluntary participation, *J. Management Educ.* **37**, 305 (2013).
- [12] N. A. Lewis, D. Sekaquaptewa, and L. A. Meadows, Modeling gender counter-stereotypic group behavior: A brief video intervention reduces participation gender gaps on STEM teams, *Soc. Psychol. Educ.*, **22**, 557 (2019).
- [13] J. P. Adams, G. Brissenden, R. S. Lindell, T. F. Slater, and J. Wallace, Observations of student behavior in collaborative learning groups, *Astron. Educ. Rev.* **1**, 25 (2002), <https://ui.adsabs.harvard.edu/abs/2002AEdRv...1a..25A>.
- [14] S. L. Eddy, S. E. Brownell, and M. P. Wenderoth, Gender gaps in achievement and participation in multiple introductory biology classrooms, *CBE Life Sci. Educ.* **13**, 478 (2014).
- [15] D. Donovan, G. Connell, and D. Grunspan, student learning outcomes and attitudes using three methods of group formation in a nonmajors biology class, *CBE Life Sci. Educ.* **17**, ar67 (2018).
- [16] P. Heller, R. Keith, and S. Anderson, Teaching problem solving through cooperative grouping. Part 1: Group versus individual problem solving, *Am. J. Phys.* **60**, 627 (1992).
- [17] P. Heller and M. Hollabaugh, Teaching problem solving through cooperative grouping. Part 2: Designing problems and structuring groups, *Am. J. Phys.* **60**, 637 (1992).
- [18] N. G. Holmes and Z. Yasemin Kalender, Preliminary evidence for available roles in mixed-gender and all-women lab groups, [arXiv:2007.14833](https://arxiv.org/abs/2007.14833).
- [19] E. M. Stump, M. Dew, S. Jeon, and N. G. Holmes, Taking on a manager role can support women's physics lab identity development, *Phys. Rev. Phys. Educ. Res.* **19**, 010107 (2023).
- [20] M. Sundstrom, D. G. Wu, C. Walsh, A. B. Heim, and N. G. Holmes, Examining the effects of lab instruction and gender composition on intergroup interaction networks in introductory physics labs, *Phys. Rev. Phys. Educ. Res.* **18**, 010102 (2022).
- [21] M. M. Camacho and S. M. Lord, "Microaggressions" in engineering education: Climate for Asian, Latina and White women, in *Proceedings of the Frontiers in Education Conference 2011* (2011), pp. S3H-1–S3H-6.
- [22] J. Premo, B. N. Wyatt, M. Horn, and H. Wilson-Ashworth, Which group dynamics matter: Social predictors of student achievement in team-based undergraduate science classrooms, *CBE Life Sci. Educ.* **21**, ar51 (2022).
- [23] S. L. Eddy, S. E. Brownell, P. Thummaphan, M.-C. Lan, and M. P. Wenderoth, Caution, student experience may vary: Social identities impact a student's experience in peer discussions, *CBE Life Sci. Educ.* **14**, ar45 (2015).
- [24] D. Z. Grunspan, S. L. Eddy, S. E. Brownell, B. L. Wiggins, A. J. Crowe, and S. M. Goodreau, Males under-estimate academic performance of their female peers in undergraduate biology classrooms, *PLoS One* **11**, e0148405 (2016).
- [25] A. N. Pell, Fixing the leaky pipeline: Women scientists in academia, *J. Animal Sci.* **74**, 2843 (1996).
- [26] T. A. Greenfield, Gender- and grade-level differences in science interest and participation, *Sci. Educ.* **81**, 259 (1997).
- [27] D. T. Burkam, V. E. Lee, and B. A. Smerdon, Gender and science learning early in high school: Subject matter and laboratory experiences, *Am. Educ. Res. J.* **34**, 297 (1997).
- [28] E. M. Smith and N. G. Holmes, Best practice for instructional labs, *Nat. Phys.* **17**, 662 (2021).
- [29] N. G. Holmes and C. E. Wieman, Introductory physics labs: We can do better, *Phys. Today* **71**, No. 1, 38 (2018).
- [30] P. Tough, *The Inequality Machine: How College Divides Us* (Mariner Books, Houghton Mifflin Harcourt, Boston, 2021).
- [31] The University of Texas at Austin, Facts & figures (2022), <https://www.utexas.edu/about/facts-and-figures>, Last accessed on 04-30-2023.
- [32] N. G. Holmes, G. Heath, K. Hubenig, S. Jeon, Z. Y. Kalender, E. Stump, and E. C. Sayre, Evaluating the role of student preference in physics lab group equity, *Phys. Rev. Phys. Educ. Res.* **18**, 010106 (2022).
- [33] S. L. Eddy and K. A. Hogan, Getting under the hood: How and for whom does increasing course structure work?, *CBE Life Sci. Educ.* **13**, 453 (2014).
- [34] S. E. Brownell, M. J. Kloser, T. Fukami, and R. J. Shavelson, Context matters: Volunteer bias, small sample size, and the value of comparison groups in the assessment

- of research-based undergraduate introductory biology lab courses, *J. Microbiol. Biol. Educ.* **14**, 176 (2013).
- [35] J. Piaget, *The Origins of Intelligence in Children* (International Universities Press, New York, 1952), Vol. 8.
- [36] B. Skinner, *The Technology of Teaching* (Appleton-Century-Crofts, 1968).
- [37] K. Sawyer, A call to action: The challenges of creative teaching and learning, *Teach. Coll. Rec.* **117**, 1 (2015).
- [38] L. C. McDermott and U. of Washington Physics Education Group, *Physics by Inquiry: Volume 1*, Physics by Inquiry Vol. 1 (Wiley, New York, 1996).
- [39] E. F. Redish, J. M. Saul, and R. N. Steinberg, On the effectiveness of active-engagement microcomputer-based laboratories, *Am. J. Phys.* **65**, 45 (1997).
- [40] E. Mazur, *Peer Instruction: A User's Manual*, Series in Educational Innovation (Prentice Hall, New York, 1997), p. 253.
- [41] M. D. Ginsburg-Block, C. A. Rohrbeck, and J. W. Fantuzzo, A meta-analytic review of social, self-concept, and behavioral outcomes of peer-assisted learning, *J. Educ. Psychol.* **98**, 732 (2006).
- [42] L. S. Vygotsky, *Mind in Society: Development of Higher Psychological Processes* (Harvard University Press, Cambridge, MA, 1978).
- [43] M. Minow, Equality vs. equity, *Am. J. Law Equality* **1**, 167 (2021).
- [44] J. M. Duncan-Andrade, *Equality or Equity: Toward a Model of Community-Responsive Education* (Harvard Education Press, Cambridge, MA, 2022).
- [45] S. J. Spencer, C. M. Steele, and D. M. Quinn, Stereotype threat and women's math performance, *J. Exp. Soc. Psychol.* **35**, 4 (1999).
- [46] J. Ehrlinger and D. Dunning, How chronic self-views influence (and potentially mislead) estimates of performance, *J. Pers. Soc. Psychol.* **84**, 5 (2003).
- [47] N. K. Campbell and G. Hackett, The effects of mathematics task performance on math self-efficacy and task interest, *J. Vocat. Behav.* **28**, 149 (1986).
- [48] E. M. Marshman, Z. Y. Kalender, T. Nokes-Malach, C. Schunn, and C. Singh, Female students with A's have similar physics self-efficacy as male students with C's in introductory courses: A cause for alarm?, *Phys. Rev. Phys. Educ. Res.* **14**, 020123 (2018).
- [49] E. A. Linnenbrink and P. R. Pintrich, The role of self-efficacy beliefs in student engagement and learning in the classroom, *Read. Writ. Q.* **19**, 119 (2003).
- [50] A. M. York, A. Fink, S. M. Stoen, E. M. Walck-Shannon, C. M. Wally, J. Luo, J. D. Young, and R. F. Frey, Gender inequity in individual participation within physics and science, technology, engineering, and math courses, *Phys. Rev. Phys. Educ. Res.* **17**, 020140 (2021).
- [51] N. G. Holmes, I. Roll, and D. A. Bonn, Participating in the physics lab: Does gender matter?, *arXiv:1905.03331*.
- [52] A. L. Graves, E. Hoshino-Browne, and K. P. Lui, Swimming against the tide: Gender bias in the physics classroom, *J. Women Minorities Sci. Eng.* **23**, 15 (2017).
- [53] Y. Copur-Gencturk, J. R. Cimpian, S. T. Lubienski, and I. Thacker, Teachers' bias against the mathematical ability of female, black, and hispanic students, *Educ. Res.* **49**, 30 (2020).
- [54] M. Sundstrom, A. B. Heim, B. Park, and N. G. Holmes, Introductory physics students' recognition of strong peers: Gender and racial or ethnic bias differ by course level and context, *Phys. Rev. Phys. Educ. Res.* **18**, 020148 (2022).
- [55] N. Davidson, *Pioneering Perspectives in Cooperative Learning: Theory, Research, and Classroom Practice for Diverse Approaches to CL* (Routledge, London, 2021).
- [56] N. Shah and C. M. Lewis, Amplifying and attenuating inequity in collaborative learning: Toward an analytical framework, *Cognit. Instr.* **37**, 423 (2019).
- [57] N. G. Holmes, Structured quantitative inquiry labs: Developing critical thinking in the introductory physics laboratory, Ph.D. thesis, University of British Columbia, 2014.
- [58] N. G. Holmes, J. Ives, and D. Bonn, The impact of targeting scientific reasoning on student attitudes about experimental physics, presented at PER Conf. 2014, Minneapolis, MN, [10.1119/perc.2014.pr.026](https://doi.org/10.1119/perc.2014.pr.026).
- [59] N. G. Holmes and D. A. Bonn, Quantitative comparisons to promote inquiry in the introductory physics lab, *Phys. Teach.* **53**, 352 (2015).
- [60] E. Etkina, D. T. Brookes, and G. Planinsic, *Investigative Science Learning Environment, 2053-2571* (Morgan & Claypool Publishers, San Rafael, CA, 2019), <https://doi.org/10.1088/2053-2571/ab3ebd>.
- [61] A. Lark, Implementation of scientific community laboratories and their effect on student conceptual learning, attitudes, and understanding of uncertainty, Ph.D. thesis, University of Toledo, Ohio, 2014.
- [62] J. Day, N. G. Holmes, I. Roll, and D. Bonn, Finding evidence of transfer with invention activities: Teaching the concept of weighted average, presented at PER Conf. 2013, Portland, OR, [10.1119/perc.2013.pr.017](https://doi.org/10.1119/perc.2013.pr.017).
- [63] The University of Texas at Austin, *Fast Facts About UT's First-Generation Student Community* (The University of Texas, Austin, 2021).
- [64] United States Department of Education, Federal Pell Grants (2023).
- [65] The University of Texas at Austin, *University of Texas Support for Low-Income Students Increases Amid Economic Challenges of the Pandemic* (The University of Texas, Austin, 2020).
- [66] The University of Texas at Austin, *UT Joins Alliance of Hispanic Serving Research Universities With 19 Other Schools* (The University of Texas, Austin, 2022).
- [67] K. Mathews, J. Phelan, N. A. Jones, S. Konya, R. Marks, B. M. Pratt, J. Coombs, and M. Bentley, 2015 national content test: Race and ethnicity analysis report, U.S. Census Bureau, Washington, DC, Technical Report, 2017.
- [68] National Center for Education Statistics, Characteristics of children's families, U.S. Department of Education, Institute of Education Sciences, Washington, DC, Technical Report No. NCES 2022144, 2022, <https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2022144>.
- [69] D. W. Johnson and R. T. Johnson, *Learning Together and Alone: Cooperative, Competitive, and Individualistic Learning* (Allyn and Bacon, Boston, 1994).
- [70] S. V. Rosser, Group work in science, engineering, and mathematics: Consequences of ignoring gender and race, *Coll. Teach.* **46**, 82 (1998).

- [71] C. Walsh, H. Lewandowski, and N. Holmes, Skills-focused lab instruction improves critical thinking skills and experimentation views for all students, *Phys. Rev. Phys. Educ. Res.* **18**, 010128 (2022).
- [72] Y. Chang and P. Brickman, When group work doesn't work: Insights from students, *CBE Life Sci. Educ.* **17**, ar52 (2018).
- [73] D. Doucette and C. Singh, Share it, don't split it: Can equitable group work improve student outcomes?, *Phys. Teach.* **60**, 166 (2022).
- [74] L. Aguilar, G. Walton, and C. Wieman, Psychological insights for improved physics teaching, *Phys. Today* **67**, No. 5, 43 (2014).
- [75] D. S. Yeager and G. M. Walton, Social-psychological interventions in education: They're not magic, *Rev. Educ. Res.* **81**, 267 (2011).
- [76] M. Dew, A. Phillips, S. Karunwi, A. Baksh, E. M. Stump, and N. G. Holmes, So unfair it's fair: Equipment handling in remote versus in-person introductory physics labs, presented at PER Conf. 2022, Grand Rapids, MI, 10.1119/perc.2022.pr.Dew.
- [77] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.20.010121> for the individual reflection assignment, partner agreement form, and partner reflection form.
- [78] J. Favela and F. Pena-Mora, An experience in collaborative software engineering education, *IEEE Softw.* **18**, 47 (2001).
- [79] M. L. Pertegal-Felices, A. Fuster-Guilló, M. L. Rico-Soliveres, J. Azorín-López, and A. Jimeno-Morenilla, Practical method of improving the teamwork of engineering students using team contracts to minimize conflict situations, *IEEE Access* **7**, 65083 (2019).
- [80] S. Brannen, D. Beauchamp, N. Cartwright, D. Liddle, J. Tishinsky, G. Newton, and J. Monk, Effectiveness of group work contracts to facilitate collaborative group learning and reduce anxiety in traditional face-to-face lecture and online distance education course formats, *Int. J. Scholarship Teach. Learn.* **15**, 5 (2021).
- [81] E. J. Theobald, M. Aikens, S. Eddy, and H. Jordt, Beyond linear regression: A reference for analyzing common data types in discipline based education research, *Phys. Rev. Phys. Educ. Res.* **15**, 020110 (2019).
- [82] J. L. Fleiss, Measuring nominal scale agreement among many raters, *Psychol. Bull.* **76**, 378 (1971).
- [83] J. L. Fleiss, B. Levin, and M. C. Paik, *Statistical Methods for Rates and Proportions*, 3rd ed. (John Wiley, New York, 2003).
- [84] B. Van Dusen and J. Nissen, Modernizing use of regression models in physics education research: A review of hierarchical linear modeling, *Phys. Rev. Phys. Educ. Res.* **15**, 020108 (2019).
- [85] L. Zhang, S. Huang, and Y. Peng, Collaboration in integrated project delivery: The effects of trust and formal contracts, *Engin. Management J.* **30**, 262 (2018).
- [86] F. L. Jeffries and R. Reed, Trust and adaptation in relational contracting, *Acad. Manag. Rev.* **25**, 873 (2000).
- [87] F. Hoffmann and P. Oreopoulos, A professor like me: The influence of instructor gender on college achievement, *J. Health Hum Resour. Adm.* **44**, 479 (2009), <http://www.jstor.org/stable/20648905>.
- [88] S. M. Solanki and D. Xu, Looking beyond academic performance: The influence of instructor gender on student motivation in STEM fields, *Am. Educ. Res. J.* **55**, 801 (2018).
- [89] M. Dew, J. Perry, L. Ford, W. Bassichis, and T. Erukhimova, Gendered performance differences in introductory physics: A study from a large land-grant university, *Phys. Rev. Phys. Educ. Res.* **17**, 010106 (2021).
- [90] A. Ozmetin, M. Dew, T. Erukhimova, and J. Perry, Does instructor gender matter for student performance in introductory physics?, presented at PER Conf. 2021, virtual conference, 10.1119/perc.2021.pr.Ozmetin.
- [91] J. Pulgar, D. Ramírez, A. Umanzor, C. Candia, and I. Sánchez, Long-term collaboration with strong friendship ties improves academic performance in remote and hybrid teaching modalities in high school physics, *Phys. Rev. Phys. Educ. Res.* **18**, 010146 (2022).
- [92] T. Lumley, P. Diehr, S. Emerson, and L. Chen, The importance of the normality assumption in large public health data sets, *Annu. Rev. Public Health* **23**, 151 (2002).
- [93] D. Bates, M. Mächler, B. Bolker, and S. Walker, Fitting linear mixed-effects models using lme4, *J. Stat. Softw.* **67**, 1 (2015).
- [94] R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2022).
- [95] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, 4th ed. (Springer, New York, 2002), ISBN 0-387-95457-0.
- [96] J. Fox and S. Weisberg, Visualizing fit and lack of fit in complex regression models with predictor effect plots and partial residuals, *J. Stat. Softw.* **87**, 1 (2018).
- [97] R. V. Lenth, emmeans: Estimated Marginal Means, aka Least-Squares Means (2023), R package version 1.8.7.