

Share, rotate, split: The effects of group work role distributions on student outcomes

Jacob Feinleib¹, Matthew Dew¹, and N. G. Holmes^{1*}

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



(Received 17 December 2025; accepted 24 February 2026; published 26 March 2026)

Education literature recommends many different strategies for structuring student group work in labs. Many of these strategies, however, have not been sufficiently evaluated for their effects on student outcomes. One prior study suggested that sharing roles, rather than splitting roles, in lab groups can boost students' physics interest and self-efficacy. Here, we expand upon this literature by evaluating the effects of a broader range of role distributions across several student outcomes in a large sample at two different institutions. We developed a survey item to probe the ways in which students distribute their roles in lab groups. The item asks for the percent of time in lab that students spent working together on lab roles (sharing), working alone on roles but rotating each session (rotating), and working alone in the same role throughout the semester (splitting). We employed hierarchical linear modeling to measure the effects of these role distributions on student critical thinking, self-efficacy, perceived agency, sense of belonging, and recognition based on survey items specific to physics lab contexts. We found that role distributions did not differentially impact student critical thinking. We also found that sharing roles tended to have a positive impact on student attitudes; splitting had a negative effect on attitudes; and rotating fell in between. Statistical significance varied across these attitudinal outcomes. Our findings invite further research and controlled studies to better understand the apparent benefits of sharing, rotating, and splitting roles in introductory physics labs.

DOI: [10.1103/1n3c-v2fj](https://doi.org/10.1103/1n3c-v2fj)

I. INTRODUCTION

Group work is utilized in many physics lab classes, especially at the introductory level. Students typically work in groups of two to three [1] to collect, analyze, and present data. While working in lab groups, students can assume a number of roles, for example, using equipment, taking notes, analyzing data, and managing the group [2]. Each of these roles involves performing a number of hands-on tasks, such as assembling a circuit for the equipment role or writing code for the analysis role. In this study, we focus on the roles that encompass performing hands-on tasks. Our prior work has investigated what roles students prefer to take on in lab [3,4] and how students spend their class time on these roles [2,4,5]. Little research has been done, however, to understand how students distribute these roles in group work.

One study sought to evaluate how different group work role distributions affect student outcomes. In an introductory physics lab, students reported a preference for a “fair split” between group members [6]. Students who reported

being in groups where members “participated equally,” however, were more likely to report that their interactions with their peers increased their physics interest and self-efficacy. The authors argued that although students might prefer to divide up work so that members specialize in a role (split), they benefit from each member working on the roles together (share); that is, when it comes to distribution of roles, “share it, don’t split it.”

In addition to groups sharing or splitting roles, students can also take turns working on each role (rotating). For example, with a group of three students A, B, and C, these role distributions would look as follows:

- **Share:** Students A, B, and C all focus on collecting data with equipment at the same time. While only one person at a time may be actively using the equipment, the other two students are actively discussing the data collection and interacting with the student taking data.
- **Rotate:** Student A uses equipment, student B analyzes data, and student C takes notes. In a subsequent lab session, student A analyzes data, student B takes notes, and student C uses equipment.
- **Split:** Throughout all their lab sessions together across the semester, student A uses equipment, student B analyzes data, and student C takes notes.

Students may choose to use one or a combination of these role distributions across the semester. Unless otherwise instructed, students typically do not discuss how they distribute their roles [2] and report that role distribution

*Contact author: ngholmes@cornell.edu

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

occurs “naturally” [3]. When asked what they *prefer*, one study found that students preferred to share roles [3]; another found students similarly preferred to share *or* split roles [4]; while another indicated that students preferred a “fair split,” which could mean sharing, rotating, or splitting [6].

When groups opt to rotate or split roles, students work on tasks individually. The risk with splitting roles is that individual students may not gain experience with particular tasks. Some researchers recommend explicitly assigning students to roles and rotating the assignment, because it ensures they gain experience with multiple roles [7] and enables students to take on roles they may not otherwise take on [8]. Other researchers argue against explicit role assignment, however, because it does not enable students to learn how to work as a group [4] or because students tend to disregard role assignments [9]. We imagine an additional risk is that students may react negatively to being assigned a role they do not want or are not good at, potentially impacting their self-efficacy or sense of belonging.

While studies have argued for different recommendations for role distributions [10–14], few base these recommendations on empirical results. Additionally, different forms of role distribution may be more or less useful for different student outcomes, whether that is learning, attitudes, or equity. For the purposes of student learning, there are theoretical reasons to expect that certain role distributions are more productive.

Here, we leverage the ICAP (Interactive, Constructive, Active, and Passive) framework, which argues that students learn better at higher levels of engagement, where engagement spans four different levels [15]. From lowest to highest engagement, these are passive, active, constructive, and interactive. In the case of students learning about measurement uncertainty in experimental physics, the following are examples at each level:

- Passive: A student *receives* information from a lecture about the best ways to minimize uncertainty for an experimental setup.
- Active: A student *manipulates* an experimental setup following instructions on the best ways to minimize uncertainty.
- Constructive: A student *generates* ideas about different ways to minimize uncertainty and tests them on an experimental setup.
- Interactive: A student *dialogues* with another student to generate and test ways to reduce uncertainty for an experimental setup.

For learning to think critically about experiments, therefore, we expect that the dialogue about the experiments (interactive) is more important for student learning than simply carrying out the tasks (active). For example, discussing the best ways to set up the equipment and collect data in order to minimize uncertainties should be more important for students’ critical thinking than carrying out the resulting procedures. Thus, we may expect that

students sharing roles most effectively motivates constructive dialogue, as students need to make decisions together about how to set up the equipment and collect data, for example. Rotating roles may next most effectively motivate constructive dialogue because, while each student is responsible for a role in a given week, other students have common proficiency with the role to support constructive dialogue. Splitting roles may least effectively motivate constructive dialogue, as each student has a unique specialty that they bring to the discussion, meaning that students may be more likely to take the recommendations of the “expert,” rather than engage in constructive dialogue.

We see no clear ordering, however, of group work role distributions for student attitudes. Each role distribution likely provides unique benefits dependent upon the attitude measure. For example, students who split roles may develop a sense of expertise (or self-efficacy) for the role they take on, while students who rotate roles may feel capable of doing each individual role. Alternatively, students who share roles may feel more socially integrated (develop a sense of belonging or recognition) because they have to interact with their team members throughout more stages of the lab.

Prior work suggests that sharing roles is more productive than splitting when it comes to students’ physics interest and self-efficacy development [6]. The study, however, has several limitations that motivate further investigation. First, rather than measuring students’ interest and self-efficacy at the start and end of the semester, the survey items asked students if their interactions with peers increased their interest and self-efficacy. Second, students were asked only a binary question about their role distribution—whether or not both they and their partner “participated equally,” which the authors interpreted as sharing roles. It is possible that students split or rotated roles in a way they believed was equal participation; the study did not comment on students’ interpretations from think-aloud interviews. Finally, the binary question does not elucidate differences between rotating and splitting roles, which prior work has noted students perceived differently [3,4].

Given the compelling claim, “share it, don’t split it,” this study invites further research. While the study investigated whether students found that peer interactions positively affected their physics interest and self-efficacy, it remains an open question whether their learning or attitudes were impacted by their group work role distribution. Importantly, do student attitudes—rather than students’ perceptions of how *peer interactions affected their attitudes*—improve? Additionally, the study looked at “conceptual guided-inquiry” labs at a single institution. Here we investigate a different context: skills-based, open-inquiry labs at two different institutions. Altogether, we seek to expand the prior investigation by evaluating more student outcome constructs across a larger population and additional instructional contexts.

In this study, we test the claim of Doucette *et al.* [6] by evaluating the effect of different group work role distributions on a range of student outcomes. Specifically, our research questions are as follows:

1. How do students report distributing their group work roles in introductory physics labs?
2. Which group work role distributions lead to better student outcomes in skills-based, open-inquiry physics labs?

II. METHODS

We detail our methodology in this section. First, in Sec. II A, we discuss our course context and student population. In Sec. II B, we compare the group work instruction across courses. In Sec. II C, we detail the measures we use for student outcomes. In Sec. II D, we discuss the survey items we developed to quantify group work role distributions. In Sec. II E, we explain how we used multiple imputation to account for missing data. Finally, in Sec. II F, we share our methodology for addressing our research questions via hierarchical linear modeling.

A. Course context and participants

Our study focuses on three introductory physics lab courses: two at institution A, a public university in the Southern U.S., and one at institution B, a private university in the Northeast U.S. These three lab courses use the Structured Quantitative Inquiry framework [16]. They are considered “open-inquiry labs” [17]: in most lab activities, students are provided with a problem or question and some theory or background, but have the opportunity to design their own procedures, choose how to analyze their data, and draw conclusions supported by their data. These labs also focus on developing experimentation skills, such as designing experiments and analyzing data, and attitudes, such as epistemological understandings of measurement (see Ref. [18] for a description of the learning goals and how activities were designed to meet them). All courses had learning goals related to group work, for example, that students should be able to brainstorm with their group to construct a diverse set of ideas when making decisions and share experimentation responsibility with other group members.

We collected survey data across two semesters. The two courses at institution A (courses A1 and A2) each had the same instructors for both semesters, whereas the course at institution B (course B) had a different instructor each semester.

Course A1 covered mechanics topics while course A2 covered electricity and magnetism, optics, and modern physics topics. Both labs were one credit hour and were corequisites to the introductory physics lecture courses. Students enrolled in these labs were in one of three lecture tracks: algebra-based physics, calculus-based physics for life science majors, and calculus-based physics for engineering

majors. The labs were typically taken earlier in student degree sequences by engineering majors and later by life science majors and those enrolled in the algebra-based sequence. See Table I for more detailed course demographics. Course A1 concluded with a lab practical exam, while course A2 concluded with a project-based lab where groups created and tested their own research questions (a level of authentic inquiry, based on Ref. [17]).

Course B was a one credit-hour lab course covering mechanics and electricity and magnetism topics and concluded with a project-based lab similar to course A2. This course was primarily taken by freshmen and sophomores in physics and engineering.

In courses A1 and A2, students had a three-hour lab session in which to collect data and had to submit their lab notes and analysis by the start of the following lab session. In course B, students completed their labs entirely within the two-hour lab session.

B. Group work instruction

All three courses involved extensive group work and a majority of each student’s final grade was determined by group-level assignments. Additionally, working collaboratively was an explicit learning objective in each course. In all three courses, students self-selected into groups of three. In situations where a group of three could not be made, students in courses A1 and A2 formed groups of two, while students in course B formed groups of four. Students changed lab groups twice throughout the semester, every three to four lab sessions.

In the first semester of courses A1 and A2, students completed multiple group work reflection assignments as detailed in a prior publication (available in the Supplemental Material of [4]). Early in the semester, students completed an individual reflection that asked them to consider their group work values (such as whether they preferred having everyone share tasks equally, a fair division of labor, or effective communication) and then to consider how they would prefer to distribute tasks, such as to get a chance to work on each task equally or work more regularly on one or two that are most comfortable to them. Students also completed “partner agreements” [4] each time they formed a new group. These partner agreements asked students to introduce themselves to each other (share preferred names and pronouns and find something they all had in common), discuss their goals for taking the course, develop a collective goal for the group, and consider multiple ways to complete their lab as a group. The instructions described the types of tasks students may need to carry out and asked students to decide as a group how they would distribute the tasks. While the prompt did not preference one type of group role distribution over another, it did point out that the tasks with which students engage would be the ones they learn most about and grow the most in their ability to perform. In sessions where

TABLE I. Course-level demographics by semester and course collected through presurveys and postsurveys. If a student's demographic response differed on the postsurvey from the presurvey, we used the postsurvey answer.

Student-level variables	Semester 1			Semester 2		
	A1	A2	B	A1	A2	B
All	815	665	618	864	560	380
Gender						
Woman	462	375	284	543	308	164
Man	336	268	316	295	230	196
Nonbinary or other	13	9	9	13	12	9
Did not disclose	4	13	9	13	10	11
Race or ethnicity						
American Indian or Alaska Native	7	5	3	10	5	1
Asian	277	257	275	346	215	239
Black or African American	48	35	64	54	36	13
Hispanic or Latino	244	169	72	244	146	22
Middle Eastern or North African	31	20	15	20	12	9
Native Hawaiian or Other Pacific Islander	3	2	3	6	0	1
White	282	230	221	255	198	113
Some other race or ethnicity	1	2	10	4	0	1
Did not disclose	4	13	9	13	10	11
Class standing						
Freshman	396	52	590	49	8	301
Sophomore	278	101	20	189	162	70
Junior	106	402	5	531	255	7
Senior	31	102	3	86	130	2
Other	0	4	0	3	2	0
Did not disclose	4	13	9	13	10	11
Major						
Physics, astronomy, or engineering physics	105	25	39	32	42	72
Engineering	259	82	488	49	135	246
Life science or biology	276	394	9	520	259	10
Other physical science	33	41	44	42	31	39
Other	0	4	0	3	2	0
Did not disclose	4	13	9	13	10	11
Parent's highest level of education						
Did not complete high school	47	25	23	39	20	14
High school/GED	139	97	103	132	80	48
Some college (but did not complete college)	64	45	22	58	43	10
Associate's degree (2-year degree)	22	34	10	32	10	8
Bachelor's degree	221	139	96	227	149	59
Master's degree	180	150	191	216	148	106
Advanced graduate degree	118	140	138	125	89	112
Not sure	3	8	12	9	4	5
Did not disclose	21	27	23	26	17	18

students did not form a new group, students completed “partner reflections” [4] as a group. These reflections asked students to reflect on ways they worked well as a group, problems they had working together, ways they could improve, and if, during the previous session, everyone worked on aspects of the lab they wanted to and if they wished to change things.

The second semester of courses A1 and A2 had very similar group work reflection opportunities. As in the first semester, students completed an individual reflection early in the lab course, and students completed partner

agreements every time they changed groups. Instead of completing partner reflections as a group every session they did not form a new group, however, students individually completed reflections once for each group they formed (3 times across the semester). This reflection was similar to the partner reflection given in the first semester, albeit the questions were slightly adjusted for individual completion and asked students to draw comparisons to elicit more substantive responses. Students were asked to describe what each group member did, if they thought everyone got to do what they wanted, how their group worked that week

compared to the previous week, and what they could do next time to work better as a group.

The first semester of course B did not involve partner agreements. Students individually completed one partner reflection a few weeks into the semester that was the same as the one given in the second semester of courses A1 and A2. The second semester of course B was similar to the second semesters of courses A1 and A2. The week-to-week delivery of assignments varied according to minor variations in the lab schedule.

In the courses at institution A, students completed their labs as a group and submitted their lab notes by the next lab period, with one submission per group. At institution B, in semester 1, students completed their lab notes as a group but then submitted individual lab reports within one week of the lab. In semester 2, students completed their labs as a group and submitted their lab notes by the end of their period, with one submission per group.

We gathered data through surveys sent to students at the beginning and end of each semester. Students were incentivized to complete the surveys with a small fraction of their grade being determined by completion of the survey (independent of student consent to participate in research). The survey was assigned to be completed outside of class. Response rates by course are shown in Table II. We used multiple imputation for missing data, as detailed in Sec. II E.

C. Outcome constructs

We measured student outcomes by testing critical thinking skills and student attitudes toward experimental physics, namely self-efficacy, perceived agency, belonging, and recognition.

1. Development of the outcome constructs

Critical thinking was measured using the Physics Lab Inventory of Critical thinking (PLIC), which has been

TABLE II. Number of participants by semester and course. To be counted as “finished,” students had to meet the following criteria: the student completed the survey; the student consented to participate in research; the student was at least 18 years of age; the student spent at least 30 s on any page; and the student provided some sort of identification (i.e., name and student ID). Data were imputed according to Sec. II E.

	Semester 1			Semester 2		
	A1	A2	B	A1	A2	B
Finished presurvey	760	606	577	773	475	362
Finished postsurvey	371	374	469	563	323	283
Finished both	316	315	428	472	238	265
Imputed presurvey	55	59	41	91	85	18
Imputed postsurvey	444	291	149	301	237	97
Total after imputation	815	665	618	864	560	380

previously evaluated for reliability and validity [19]. An additional study [20] motivated us to remove two items from the PLIC related to evaluating methods. The study found that the two items did not sufficiently engage students’ critical thinking as compared to the final evaluating methods item. The study also found that removing those items did not affect students’ reasoning on the final evaluating methods item.

The development of the attitudes items was partially reported in Refs. [21,22]. Items were either taken from or developed based on existing survey instruments or literature about each construct. Two self-efficacy items came from Refs. [23,24]. The other self-efficacy items were inspired by the task-specific items in Ref. [23], but developed to relate to specific tasks students carry out in introductory physics labs. Three of the perceived agency items came from Ref. [25], again adapted to be in the context of physics labs. One perceived agency item was created to reflect Scardamalia and Bereiter’s [26] definition of epistemic agency. These four items were previously used in Ref. [21] to study students’ sense of agency. All four belonging items came from Refs. [24,27], though contexts were adapted from “in a math setting” and “in this class” to being “in a physics lab.” Finally, all four recognition items were adapted from those in Ref. [28], though recognition from parents/relatives/friends/physics teacher was adapted to be from lab instructors/lab peers and physics instructors/physics peers outside of lab.

Researchers conducted interviews with physics, life sciences, and engineering majors enrolled in introductory physics lab courses at institutions A and B to evaluate the construct validity [29]. During the interviews, students were asked to complete the survey and think out loud while they responded to the items. They were also asked to clarify their thinking and how they were interpreting items. The interviews led to some items being reworded and some items being dropped from the survey. Interviews continued until saturation was achieved (15 total interviews). We further checked the validity of these outcome construct items using confirmatory factor analysis, as described below.¹

2. Confirmatory factor analysis assumptions

We used confirmatory factor analysis to evaluate the extent to which the survey items reflect the theoretical relationships between constructs. First, we evaluated whether the data met the assumptions necessary for confirmatory factor analysis, following recommendations from Knekta *et al.* [30].

¹Two belonging items were excluded from the analysis early in the confirmatory factor analysis due to low factor loadings. The numbers we report below are from the final confirmatory factor analysis without these items.

Combining the presemester and postsemester survey responses, we had 5904 responses.² Of these 5904 responses, we were missing at least one construct-related item for 2.9% of responses. We investigated our data for outliers and found that student responses spanned the full Likert scale for all items, although some had narrow distributions. We calculated the Mahalanobis distances [31,32] for our data to identify potential outliers. Upon inspection, we did not find any reasons (such as students selecting the same choice for all questions) to drop these potential outliers.

To evaluate factorability, we used Kaiser's measure of sampling adequacy. We obtained a measure of sampling adequacy of 0.93, well above the threshold of 0.6 [30]. We also checked the interitem correlations and found that the minimum correlation among items from the same factor was 0.43, which is above the 0.30 threshold [30].

Next, we checked the normality and linearity of our data. We assessed univariate normality with skewness and kurtosis. No item should have a skewness or kurtosis above $|2.0|$ [30]; our highest value among either of these was 1.16. We checked multivariate normality using Mardia's multivariate normality test and found statistically significant skewness ($p < 0.001$) and kurtosis ($p < 0.001$) for our data. We addressed this in our analysis by using an estimator for confirmatory factor analysis that is robust against non-normality. We checked linearity within factors by plotting—for each factor—student scores for every item against their scores on that factor's other items. We also calculated Spearman's rank-order correlation coefficient between factor items and found all coefficients were greater than 0.4 and statistically significant (< 0.001), indicating sufficient linearity [30].

Finally, we checked multicollinearity using variance inflation factors. Our largest value was 4.03, which is less than the maximum threshold of 10 [30] indicating that our data are not multicollinear.

With all these checks, our data met the assumptions necessary for confirmatory factor analysis, so long as we used a methodology robust to non-normality. The full set of figures and values used for checking our assumptions is available at Ref. [33].

3. Confirmatory factor analysis results

We performed confirmatory factor analysis with a four-factor structure: one factor each for self-efficacy, perceived agency, belonging, and recognition (Table III). We used maximum likelihood estimation with robust standard errors to account for the non-normality of our data. The chi-squared test of model fit was statistically significant ($\chi^2 = 2963.179$, d.o.f. = 203, $p < 0.001$), suggesting our data do not fit the

model. The chi-squared value typically increases as the data size increases, however, which may mean our model was unfairly rejected due to small differences [30,34]. We consider this limitation in the context of other fit indices.

We obtained a comparative fit index (CFI) of 0.956, a root mean square error of approximation (RMSEA) of 0.056, and a standardized root mean square residual (SRMR) of 0.030. All of these measures of goodness-of-fit meet the desired thresholds: CFI above 0.95, RMSEA below 0.06, and SRMR below 0.08 [30].

Our factor loadings are given in Fig. 1 and Table III (the latter with the full list of questions). All items, except two, had factor loadings above 0.7. The two items that did not meet this threshold were SE_1 (“Express my opinions when others disagree with me during the lab,” which had a loading of 0.61) and SE_4 (“Interpret graphs of my measurements,” which had a loading of 0.68). We kept these items for several reasons: the loadings are both near the threshold, SE_1 was almost identical to an item from Ref. [23], and both items were considered important to the construct of lab self-efficacy.

D. Roles distribution survey development

We iteratively developed a survey item to measure the ways in which students distributed their roles during their collaborative lab work. The goal of the item was to serve as a postsemester estimate of how students tended to distribute their roles across the full course. First, we administered a pilot open-response question on a post-survey to an introductory physics lab course at institution B. A preceding question asked students about the frequency with which they took on several hands-on and cognitive (such as decision-making) roles in the lab. The open-response roles distribution question read:

Use the space below to explain how your group split up the above roles. (i.e., did one person do a role, did multiple people work independently, or did everyone work together?)

We first read each text response ($N = 1037$) to this question and found a number of themes common across responses:

- Working on all roles together.
- Individuals taking on roles that they did not take on during the previous lab session.
- Individuals specializing in and staying with roles or doing roles with which they had experience.

These themes expand upon the share or split framework [6] to include the behavior of students working individually but switching roles between sessions, which we refer to as rotating. The themes also echo those found in an interview study conducted at institution B [3].

We further inspected responses that included a specific reason indicating why a group distributed roles in a certain

²This response count is 32 fewer responses than the sum of the first two rows in Table II because some students opted to skip all four attitude construct questions.

TABLE III. Survey items by construct. The self-efficacy items prompt reads, “Please rate how confident you are that you can do each of the following things in a physics lab course.” The answer options for the self-efficacy items ranged from “Not Confident” to “Confident” on a five-point Likert scale. The other item prompts read, “Please indicate how well you agree with the following statements” and had answer options on a five-point Likert scale ranging from “Strongly Disagree” to “Strongly Agree.” Items marked with * were reverse coded.

Construct	Item number	Item	Factor loading
Self-efficacy	SE_1	Express my opinions when others disagree with me during the lab.	0.61
	SE_2	Overcome any problems I encounter in an experiment.	0.72
	SE_3	Interpret experimental outcomes taking into account experimental uncertainty.	0.76
	SE_4	Interpret graphs of my measurements.	0.68
	SE_5	Design reliable experiments with the available equipment.	0.82
	SE_6	Make accurate predictions about experimental outcomes.	0.73
	SE_7	Comfortably take charge of the equipment in the physics lab.	0.71
	SE_8	Design an experiment that answers my research question.	0.84
	SE_9	Generate further research questions based on my observations in the lab.	0.74
	SE_10	Design an experiment in a physics lab where I can find something I did not know before.	0.78
Perceived agency	PA_1	I am in control of setting the goals for the experiments.	0.82
	PA_2	I have the freedom to design and conduct the best possible experiment to attain my goals.	0.88
	PA_3	I am in control of choosing the appropriate analysis tools to evaluate experimental outcomes.	0.86
	PA_4	I am in control of doing interesting experiments in a physics lab.	0.79
Belonging	BEL_1	I feel like an outsider in a physics lab.*	0.79
	BEL_2	When I get a poor grade on an experiment, I feel that maybe I don't belong in a physics lab.*	0.75
	BEL_3	Sometimes I worry that I do not belong in a physics lab.*	0.86
	BEL_4	When I am in a physics lab, I wish I could fade into the background and not be noticed.*	0.79
Recognition	REC_1	My lab peers see me as a physics person.	0.88
	REC_2	My lab instructors see me as a physics person.	0.88
	REC_3	My physics peers outside of lab see me as a physics person.	0.87
	REC_4	My physics instructors outside of lab see me as a physics person.	0.89

way ($N = 80$). After excluding themes that appeared with low frequency, we identified five recurring themes in student responses:

- Sharing so everyone could gain experience.
- Rotating so everyone could experience different roles.
- Splitting so people could do what they enjoyed.
- Splitting because of logistical constraints.
- Naturally falling into roles.

Students generally used the word “naturally” to mean that there was no rationale motivating their distribution of lab roles.

We converted the open-response question into a closed-response format where students indicated how and why they distributed their group work roles. The answer choices for this question were the three common types of role distribution from the open-response question: sharing, rotating, and splitting. Each choice also included a positive reason to choose it so that no option seemed undesirable. For example, one choice read, “We split roles so people could do what they enjoyed.” Here, enjoyment was the positive reason to choose splitting. The question thus read:

During your time *in the lab*, how did your group distribute the lab roles (equipment, notes, analysis)? Why?

- We all worked together on every role so that everyone could learn about and experience every role.
- We divided the roles and rotated who worked on each role so that everyone could learn about and experience every role.
- We divided the roles and kept the same roles each week so each person could do what they enjoyed or were experienced with.
- We divided the roles based on logistical constraints (i.e., data or code was on one person's computer).
- We naturally fell into roles without any discussion.
- Other.

Students who selected “Other” were prompted with a text entry box to describe their response in more detail.

In transferring this question to an online survey format, we used a constant-sum response format, where students

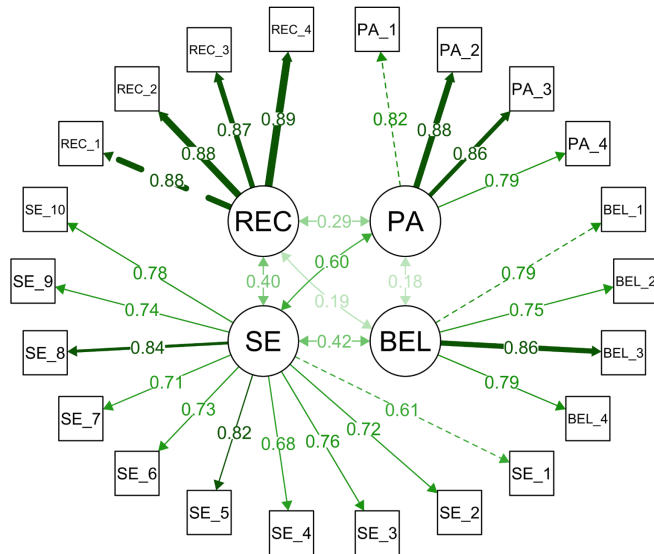


FIG. 1. Factor loadings from confirmatory factor analysis of outcome constructs. “SE” is self-efficacy, “PA” is perceived agency, “BEL” is belonging, and “REC” is recognition. Anchor items for each factor are denoted by dashed lines. Line thickness and shade denote factor-loading magnitude. Table III gives the full text for each survey item.

could allocate up to 100% of their time among the options. An allocation of 100% to an item would indicate that their group(s) distributed their lab roles in that way throughout the whole semester. An allocation of 50% to an item would indicate that their group(s) distributed their lab roles in that way about half of the time.

We conducted interviews with four students in course B to evaluate the question and answer choices. In these interviews, interviewees answered the questions about their most recent lab course. As interviewees completed the survey, we asked them to think out loud, particularly to articulate how they interpreted each question and which specific lab roles they had in mind for each question. We had three main takeaways from this set of interviews. First, the “natural” option was overly broad and could encompass both sharing and splitting roles. With the constant-sum format, we wanted to avoid redundancy. Second, we did not need two separate options for splitting roles with different reasons. Our research questions were more concerned with the actual role distribution than the student’s rationale. Third, we found that students struggled to distinguish working together cognitively with dividing or sharing the hands-on roles.

These findings led us to reduce the number of options from six to four and we reworded the options to have the positive reason be in parentheses to act as a suggestion rather than a criterion. In addition, we adjusted the question to specify that we were asking about hands-on roles as opposed to decision-making roles. The prompt and answer choices read:

During your time in the lab, how frequently did your group organize **physical, hands-on lab roles** (equipment, notes, and analysis) in the following ways? Enter a percentage of time for each option below where 0% is “Never” and 100% is “Always.” Your percentages should add up to 100%.

- We worked together on roles (for example, so that we could bounce ideas off each other for each role).
- We divided the roles and rotated who worked on each role (for example, so that everyone could learn about and experience every role).
- We divided the roles and kept the same roles each week (for example, so that each person could do what they enjoyed or were experienced with).
- Other.

Students who selected “Other” were prompted with a text entry box to describe their response in more detail.

We then interviewed six students taking course A1 or A2, with the goal of validating the survey questions with a new population of students. The interviewees generally interpreted the items as intended. The only adjustment made from these interviews was to bold the words “rotated” and “same” in the second and third answer choices, respectively. This change occurred after the second interview. The following four interviews brought us to saturation.

E. Multiple imputation

Our outcome constructs were measured through pre-semester and postsemester surveys, and the group work role distribution item was implemented only on the post-semester survey. Matching across presurveys and postsurveys led to missing data, and so we follow prior recommendations to impute data [35]. We imputed our data following the directions in Ref. [36], particularly the supplemental decision tree [37].

Our percentages of missing data for each variable in each course by semester are given in Table IV. While these missingness percentages vary across courses, prior work has found that multiple imputation is robust across various situations, such as regression models with up to 18 predictors, missingness percentages up to 50%, and sample sizes as small as $N = 50$ [38].

First, before imputing data, we evaluated whether we had access to other variables related to our missingness. The most common explanatory variable used in imputation for missing postsemester data is final grade [35]. Students with lower final grades are less likely to complete postsemester surveys [35]. We only had access to grade data for semesters 1 and 2 of course B. We did not have access to grade data for courses A1 and A2 for either semester. Thus, we checked if postsemester survey missingness was explained by presemester survey data for courses A1 and A2.

TABLE IV. Percentage of missingness for survey items on the presurveys and postsurveys by course and semester.

Survey item	Semester 1			Semester 2		
	A1 (%)	A2 (%)	B (%)	A1 (%)	A2 (%)	B (%)
Presurvey						
PLIC	6.7	8.9	6.6	10.5	15.2	4.7
Self-efficacy	9.2	10.7	8.7	12.4	17.0	5.0
Perceived agency	8.2	9.8	9.1	12.6	17.3	5.0
Belonging	7.7	10.1	8.4	12.5	16.4	5.3
Recognition	7.9	10.2	7.8	12.2	17.0	5.0
Postsurvey						
PLIC	54.5	43.8	24.1	34.8	42.3	25.5
Self-efficacy	55.2	44.4	25.1	35.8	43.2	26.1
Perceived agency	55.0	44.2	24.3	35.3	43.2	25.5
Belonging	55.2	44.2	24.6	35.5	43.0	25.5
Recognition	55.2	44.4	24.6	35.3	43.2	25.5
Share	54.5	43.8	24.1	34.8	42.3	25.5
Rotate	54.5	43.8	24.1	34.8	42.3	25.5
Split	54.5	43.8	24.1	34.8	42.3	25.5

For courses A1 and A2, we checked if presemester PLIC scores were a statistically significant predictor of postsemester survey missingness. Using a t test, we found presemester PLIC scores were a statistically significant predictor for both course A1 ($p = 0.026$) and course A2 ($p < 0.001$). This test indicated that students with higher presemester PLIC scores were more likely to complete the postsemester survey. We also ran a logistic regression for each course controlling for random effects due to the semester offering and found that presemester PLIC scores were still predictors of postsemester survey missingness for course A1 ($\beta = -0.835$, $p = 0.048$) and course A2 ($\beta = -2.492$, $p < 0.001$).

In course B, we inspected whether final grades were a statistically significant predictor of postsemester survey missingness. For final grades, we used final letter grades on a 4.3 scale, which is a typical 4.0 scale with +’s adding 0.3 and –’s subtracting 0.3 from the typical letter grade point. We ran a t test across all course B data and found that students with higher grades were more likely to complete the survey ($p < 0.001$). We similarly ran a logistic regression controlling for random effects due to the semester offering and found that grade was still a predictor of postsemester survey missingness ($\beta = -4.283$, $p < 0.001$).

For all three courses, we checked if postsemester PLIC scores predicted presemester survey missingness. Using t tests, we found that postsemester PLIC scores predicted presurvey missingness for course A1 ($p = 0.016$), course A2 ($p = 0.003$), and course B ($p = 0.002$). We also ran logistic regressions for each course controlling for random effects due to the semester offering and found that

TABLE V. Variables available for imputation for each course and semester. We only had access to student grades for course B.

Survey item	Semester 1			Semester 2		
	A1	A2	B	A1	A2	B
Presurvey						
PLIC	X	X	X	X	X	X
Self-efficacy	X	X	X	X	X	X
Perceived agency	X	X	X	X	X	X
Belonging	X	X	X	X	X	X
Recognition	X	X	X	X	X	X
Postsurvey						
PLIC	X	X	X	X	X	X
Self-efficacy	X	X	X	X	X	X
Perceived agency	X	X	X	X	X	X
Belonging	X	X	X	X	X	X
Recognition	X	X	X	X	X	X
Share	X	X	X	X	X	X
Rotate	X	X	X	X	X	X
Split						
Demographics						
Gender	X	X	X	X	X	X
Race or ethnicity	X	X	X	X	X	X
Class standing	X	X	X	X	X	X
Major	X	X	X	X	X	X
Parent’s highest level of education	X	X	X	X	X	X
Grade			X			X

postsemester PLIC scores were still predictors for presemester survey missingness for course A1 ($\beta = -1.718$, $p = 0.013$), course A2 ($\beta = -2.299$, $p = 0.001$), and course B ($\beta = -3.381$, $p < 0.001$).

These checks indicate variables that act as significant predictors for missingness across our dataset. To prepare for imputation, we standardized all our variables to range from 0 to 1. Variables that were available for imputation are given in Table V. We used the `quickpred` function from the MICE package in *R* to determine which variables should be used to impute each variable [39]. Our predictor matrix is given in Fig. 2.

We imputed all the presemester and postsemester outcome constructs (PLIC, self-efficacy, perceived agency, belonging, and recognition) as well as responses for the role distribution item. We used the imputation method `21.pmm` for the outcome constructs, which is predictive mean matching for multilevel data. Predictive mean matching uses values that exist within the dataset, so it is more robust in cases of nonlinear data and therefore more widely applicable [40].

Because the three group work role distribution choices (share, rotate, and split) must add up to 100%, we used predictive mean matching imputation for the share and rotate items and then used passive imputation for the split

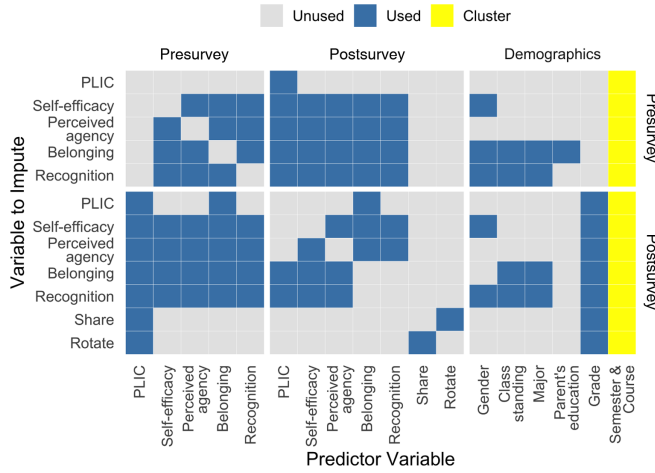


FIG. 2. Predictor matrix for multiple imputation. “Semester & Course” was used as a cluster variable for all variables. We passively imputed data for “Split” (not shown here).

item. Passive imputation simply means we imputed one value based on other values with a predefined equation. In this case, we defined Split as $1 - \text{Share} - \text{Rotate}$.³

According to prior recommendations [41,42], our number of imputations m should meet the following threshold,

$$FMI/m \leq 0.01$$

where FMI is the fraction of missing information [41,42]. We tried 5, 10, 30, and 50 imputations and found that 50 was sufficiently many imputations so that every imputed variable had a low enough FMI/m ratio. Our results use multiple imputation with $m = 50$. Table VI shows the descriptive statistics across all courses, semesters, and constructs before and after imputation.

F. Hierarchical linear modeling

We evaluated the relationships between outcome constructs and group work role distributions using hierarchical linear modeling [43,44]. Hierarchical linear modeling accounts for potential systematic differences within nested data. Consider a study of the effects of an intervention conducted during recitation on student postsemester scores in a large introductory course. The intervention, being delivered in recitation sections, could be implemented to varying degrees of effectiveness—one TA might more faithfully implement it than others, one section might be late at night, one classroom layout might be less useful for

³We also technically imputed grade, but this was required by the MICE package. To use grade for course B, MICE needs an imputed variable for grades in courses A1 and A2. We do not discuss imputation for this variable since it acts as a dummy variable for these courses.

discussions, etc. There are myriad possible effects from section-level differences that are difficult or impossible to control. Hierarchical linear modeling allows us to account for all these indeterminate effects that take place within this nested data. For an example comparing the results of hierarchical linear modeling to alternate methods, see Sec. III B of Ref. [43].

In this study, we use hierarchical levels to account for potential systematic differences between courses. For example, these differences could include the semester in which the course was offered, which may indicate whether students are in their first or second semester of college; the institution, as different institutions having different populations of students; and the course itself, as differences in the specific content covered in each course may lead to variations in student performance.

We used a two-level model to connect a student’s postsemester score on an outcome construct to their method of role distribution. We normalized the scores of the outcome constructs (i.e., to a 0–1 scale) to account for differing scales between the PLIC and attitude constructs. We clustered responses based on the unique combination of course and semester. We modeled our level 1 data as in Eq. (1)—the score for a postsemester construct, $Post$, for student i in (semester + course) j —such that

$$Post_{ij} = \beta_{0j} + \beta_{1j}Pre + \beta_{2j}RD + r_{ij} \quad (1)$$

where β_{0j} is the intercept; Pre is the i th student’s score for the presemester construct in (semester + course) j ; RD was the i th student’s reported fraction of time spent on a given role distribution type in (semester + course) j ; and r_{ij} is the residual term.

We modeled our level 2 data as in Eq. (2) such that,

$$\beta_{0j} = \gamma_{00} + u_{0j}, \quad (2)$$

where γ_{00} is the mean intercept and u_{0j} is the residual term. We check that the assumptions of hierarchical linear modeling are met in the Appendix.

We theorized, based on previous work [6], that gender may be a significant factor that could also impact our model. However, adding a gender term to the model increased both AIC and BIC for all combinations of role distribution type and outcome construct for the data before multiple imputation. For the imputed dataset, the AIC and BIC either increased or stayed approximately the same for all combinations of role distribution type and outcome construct. This indicates the gender term did not meaningfully improve the model. Thus, we did not include gender in our final model.

Beyond gender, we found insufficient theoretical motivation to include additional demographic characteristics (such as student race or ethnicity or major) as possible predictors of course outcomes. One study evaluated the

TABLE VI. Descriptive statistics (mean and standard deviation) using listwise deletion and using multiple imputation for each course and semester.

Survey item	Semester 1			Semester 2		
	A1	A2	B	A1	A2	B
Using listwise deletion						
Presurvey						
PLIC	0.39 ± 0.12	0.44 ± 0.13	0.42 ± 0.13	0.40 ± 0.12	0.45 ± 0.12	0.45 ± 0.12
Self-efficacy	0.66 ± 0.18	0.73 ± 0.18	0.70 ± 0.18	0.68 ± 0.18	0.72 ± 0.17	0.70 ± 0.15
Perceived agency	0.71 ± 0.18	0.72 ± 0.20	0.67 ± 0.19	0.71 ± 0.19	0.71 ± 0.20	0.69 ± 0.18
Belonging	0.46 ± 0.21	0.59 ± 0.23	0.57 ± 0.22	0.46 ± 0.21	0.51 ± 0.24	0.56 ± 0.20
Recognition	0.47 ± 0.17	0.47 ± 0.21	0.51 ± 0.19	0.47 ± 0.19	0.48 ± 0.20	0.54 ± 0.18
Postsurvey						
PLIC	0.45 ± 0.13	0.45 ± 0.13	0.44 ± 0.13	0.43 ± 0.13	0.46 ± 0.13	0.52 ± 0.14
Self-efficacy	0.82 ± 0.14	0.82 ± 0.16	0.84 ± 0.14	0.81 ± 0.17	0.80 ± 0.16	0.81 ± 0.14
Perceived agency	0.78 ± 0.19	0.74 ± 0.22	0.76 ± 0.19	0.74 ± 0.21	0.74 ± 0.22	0.74 ± 0.21
Belonging	0.59 ± 0.26	0.63 ± 0.25	0.67 ± 0.22	0.59 ± 0.24	0.60 ± 0.24	0.64 ± 0.23
Recognition	0.53 ± 0.21	0.52 ± 0.23	0.56 ± 0.21	0.51 ± 0.22	0.51 ± 0.23	0.56 ± 0.21
Share	0.49 ± 0.31	0.46 ± 0.29	0.42 ± 0.25	0.39 ± 0.27	0.39 ± 0.30	0.33 ± 0.21
Rotate	0.21 ± 0.22	0.16 ± 0.19	0.26 ± 0.19	0.20 ± 0.20	0.17 ± 0.21	0.22 ± 0.21
Split	0.31 ± 0.29	0.38 ± 0.30	0.32 ± 0.25	0.41 ± 0.29	0.43 ± 0.32	0.45 ± 0.27
Using multiple imputation						
Presurvey						
PLIC	0.38 ± 0.13	0.42 ± 0.13	0.41 ± 0.13	0.39 ± 0.12	0.42 ± 0.13	0.44 ± 0.13
Self-efficacy	0.67 ± 0.18	0.72 ± 0.18	0.69 ± 0.18	0.68 ± 0.18	0.72 ± 0.18	0.71 ± 0.16
Perceived agency	0.71 ± 0.19	0.72 ± 0.20	0.67 ± 0.20	0.71 ± 0.20	0.71 ± 0.19	0.70 ± 0.18
Belonging	0.46 ± 0.22	0.55 ± 0.24	0.56 ± 0.22	0.46 ± 0.21	0.51 ± 0.23	0.56 ± 0.20
Recognition	0.48 ± 0.18	0.49 ± 0.21	0.51 ± 0.20	0.47 ± 0.19	0.49 ± 0.20	0.54 ± 0.18
Postsurvey						
PLIC	0.43 ± 0.13	0.44 ± 0.13	0.44 ± 0.13	0.43 ± 0.13	0.46 ± 0.13	0.50 ± 0.14
Self-efficacy	0.81 ± 0.16	0.82 ± 0.16	0.83 ± 0.15	0.80 ± 0.17	0.82 ± 0.16	0.82 ± 0.14
Perceived agency	0.77 ± 0.20	0.75 ± 0.22	0.76 ± 0.20	0.74 ± 0.21	0.75 ± 0.22	0.75 ± 0.21
Belonging	0.58 ± 0.25	0.61 ± 0.25	0.66 ± 0.23	0.57 ± 0.25	0.60 ± 0.25	0.64 ± 0.23
Recognition	0.52 ± 0.22	0.53 ± 0.23	0.56 ± 0.22	0.51 ± 0.23	0.52 ± 0.23	0.57 ± 0.22
Share	0.48 ± 0.30	0.45 ± 0.30	0.43 ± 0.27	0.38 ± 0.26	0.37 ± 0.26	0.34 ± 0.22
Rotate	0.19 ± 0.20	0.19 ± 0.21	0.25 ± 0.20	0.21 ± 0.20	0.20 ± 0.21	0.22 ± 0.21
Split	0.33 ± 0.28	0.36 ± 0.33	0.32 ± 0.26	0.41 ± 0.28	0.43 ± 0.30	0.44 ± 0.27

effects of student major (through lecture tracks) and found no consistent trend in lab group work preferences nor equipment handling [4]. Another study evaluated the effects of race or ethnicity and found no effects on equipment handling [45]. We, therefore, excluded these other demographic variables from our models to avoid gap gazing [46] and overfitting.

III. RESULTS

A. Role distributions

The results from the role distribution survey question before multiple imputation are shown in Fig. 3. Across all courses in semester 1, we found that students reported sharing roles (green) at higher rates than other distributions (evidenced by higher frequencies of greater reported times), followed by splitting (blue). In semester 2, splitting was as

frequent as or more frequent than sharing. In both semesters, rotating (orange) was the least frequently reported role distribution (evidenced by higher peaks at the lowest reported times).

We found that across all classes, many students reported sharing roles between 25% and 50% of their time in lab. More students in courses A1 and A2 than in course B reported sharing roles for more than 50% of the time. Course B had especially few students reporting sharing roles more than 50% of the time in semester 2 compared to any other course.

We found that many students reported rotating roles very rarely. In courses A1 and A2 and semester 2 of course B, the rotating curve peaks at the lowest reported times and very few students reported rotating roles for the majority of their time in lab. In semester 1 of course B, however, the rotating curve peaks at about 25% of the reported time.

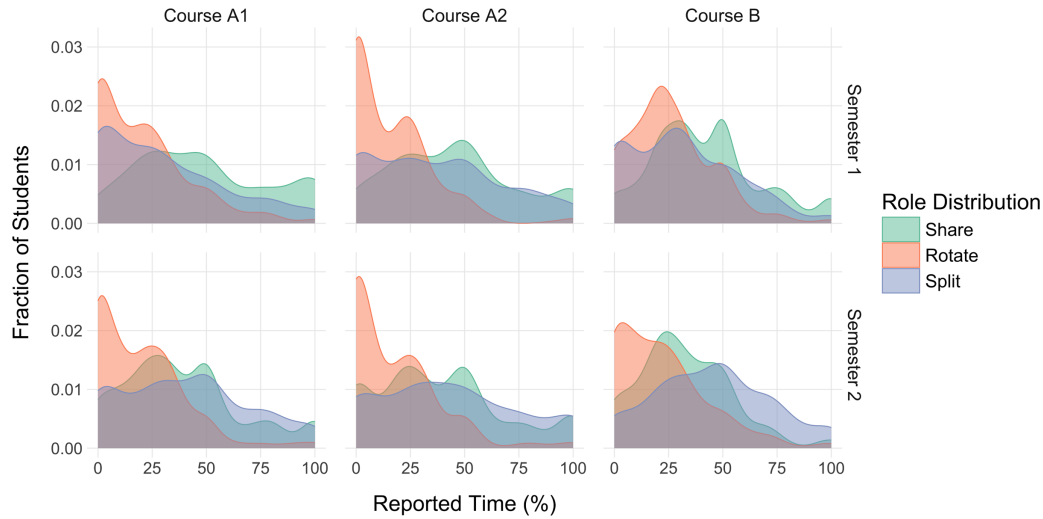


FIG. 3. Density plots of the reported percent of lab time students spent sharing, rotating, or splitting roles. Data are presented here using listwise deletion.

Our data show that the percent of time in lab that students split roles was more evenly distributed compared to other role distribution methods. In semester 1, we found that a greater fraction of students spent less than half of their time in lab splitting roles, whereas in semester 2, we found that most students reported spending about half of their time in lab splitting roles.

B. Construct results

Students' overall presurvey and postsurvey scores on the outcome constructs are shown in Fig. 4. We collapsed the data for semesters 1 and 2 for each course because the score distributions were similar. We found that students' scores increased from presurvey to postsurvey for all constructs,

though the size of the increases varied by course and construct. The attitudinal constructs tended to have skewed distributions with ceiling effects, particularly self-efficacy and perceived agency.

C. Hierarchical linear modeling results

The results of the hierarchical linear modeling are in Fig. 5 and Table VII. The figure indicates the effect size for each outcome construct, such that a positive effect means more reported time in a particular role correlates to higher postsurvey scores, controlling for presurvey scores and random effects. We found that a student's role distribution had no significant effect on their PLIC performance.



FIG. 4. Bar graphs of the distribution of student scores on the PLIC and attitude constructs. The colored region represents the interquartile range and the horizontal line represents the median score. Whiskers indicate the values at 1.5 times the interquartile range above the third quartile and below the first quartile. Dots indicate outliers defined as scores beyond the range of the whiskers (data are presented here using listwise deletion).

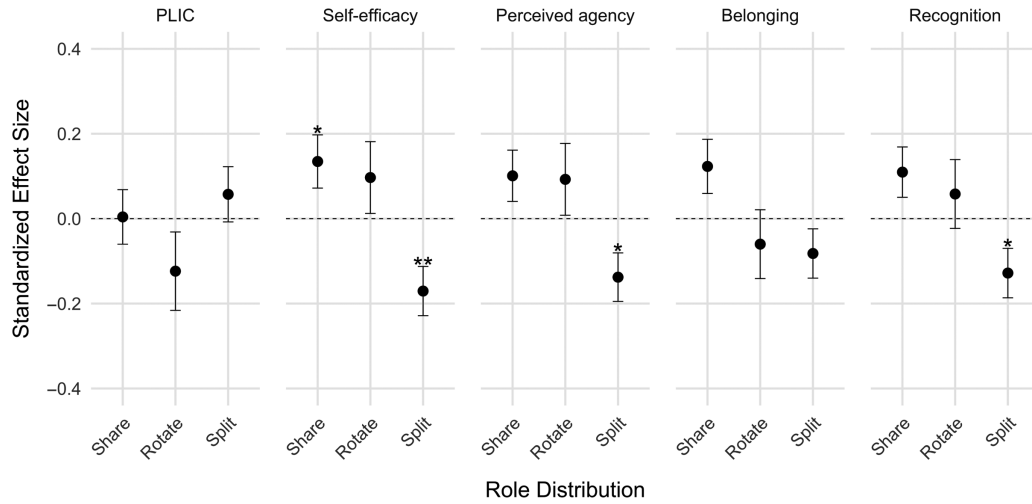


FIG. 5. Results of hierarchical linear modeling of change in PLIC and attitude scores controlling for semester and course using data after multiple imputation. A value greater than zero indicates a positive effect on score, whereas a value less than zero indicates a negative effect on score. The horizontal dashed line indicates no observable effect on score. The error bars represent the standard error of the effect size. The asterisks denote statistical significance where * indicates $p < 0.05$, and ** indicates $p < 0.01$.

Student role distributions did have an effect on attitude constructs, with similar trends across each construct, but varying statistical significance. In all cases, more time spent sharing roles corresponded to the strongest positive effect sizes (statistically significantly greater than zero only for self-efficacy). More time spent splitting roles corresponded to the strongest negative effect sizes (statistically significantly less than zero for self-efficacy, perceived agency, and recognition). The effect of rotating varied between constructs (sometimes positive and sometimes negative effect sizes), but was never statistically distinguishable from zero.

IV. DISCUSSION

In this study, we evaluated the effect of group work role distributions (whether sharing, rotating, or splitting hands-on lab roles) on several student outcomes: critical

thinking, self-efficacy, perceived agency, belonging, and recognition. Here we analyze our results through the lens of our research questions.

A. Group roles distributions

Despite different student populations, we found that students across courses and semesters reported sharing or splitting roles much more frequently than rotating roles. Students' role distributions may relate to their group work preferences. At institution A, a previous study [4] found that students preferred sharing and splitting roles, with the fewest students preferring to rotate roles, consistent with their reported role distributions here. At institution B, however, a previous study [3] found that most students—especially women—preferred sharing roles. These preferences are consistent with the data from semester 1, but not from semester 2.

TABLE VII. Results from hierarchical linear modeling for change in outcome construct scores. The table includes the standardized effect sizes, standard errors, and p values in parentheses.

Construct	Share	Rotate	Split
Using listwise deletion			
PLIC	0.0579 ± 0.0766 (0.450)	-0.2321 ± 0.1029 (0.024)	0.0645 ± 0.0733 (0.379)
Self-efficacy	0.2443 ± 0.0725 (0.001)	0.2276 ± 0.0979 (0.020)	-0.3396 ± 0.0690 (< 0.001)
Perceived agency	0.1695 ± 0.0735 (0.021)	0.2286 ± 0.0997 (0.022)	-0.2725 ± 0.0702 (< 0.001)
Belonging	0.2336 ± 0.0689 (0.001)	-0.1226 ± 0.0933 (0.189)	-0.1532 ± 0.0661 (0.021)
Recognition	0.1779 ± 0.0700 (0.011)	0.1783 ± 0.0949 (0.060)	-0.2557 ± 0.0668 (< 0.001)
Using multiple imputation			
PLIC	0.0041 ± 0.0642 (0.949)	-0.1237 ± 0.0923 (0.181)	0.0574 ± 0.0650 (0.378)
Self-efficacy	0.1347 ± 0.0626 (0.032)	0.0968 ± 0.0845 (0.253)	-0.1705 ± 0.0580 (0.003)
Perceived agency	0.1009 ± 0.0604 (0.095)	0.0926 ± 0.0846 (0.274)	-0.1378 ± 0.0571 (0.016)
Belonging	0.1231 ± 0.0638 (0.054)	-0.0600 ± 0.0810 (0.459)	-0.0820 ± 0.0580 (0.158)
Recognition	0.1096 ± 0.0593 (0.065)	0.0581 ± 0.0810 (0.474)	-0.1281 ± 0.0583 (0.028)

Alternatively, prior research suggests that student preferences do not sufficiently predict the roles that they actually take on [3,4]. An alternative explanation of the role distributions data, therefore, is the use of partner agreements [4]. Courses A1 and A2 as well as semester 2 of course B all used partner agreements and their distributions look similar in Fig. 3. Semester 1 of course B, however, did not use partner agreements and looks slightly different than the other courses; in particular, more students reported rotating roles. While we do not have data on student preferences for these semesters, we can reasonably infer that preferences are similar across semesters within each course. This suggests that partner agreements may have a larger effect on student role distributions. Future work should test this hypothesis directly.

We also note spikes in the frequency of students who reported rotating roles for 25% of their time in lab, splitting roles between 25% and 50% of their time, and sharing roles for 50% of their time. Despite our survey question offering students a continuous range of reported times, it is possible that students divided their reported time into discrete chunks, such as quarters, halves, or thirds, resulting in these spikes. It is also possible that many students were reluctant to select the most extreme values for the percentage of time spent using each type of role distribution, i.e., 0% and 100%. Thus, analyzing the data along a continuous scale could overestimate the precision of students' perceptions of how they spent their time. The distributions themselves, however, suggest a reasonably continuous distribution of student responses that are sufficient for our linear regression analyses. Thus, we do not believe this limitation affected our analysis, particularly since our models pass our assumption tests (see the Appendix and Ref. [33]).

B. Outcome constructs

We find that student scores increased from presurvey to postsurvey across all measures and in all courses. The current study is the first to report data about students' responses to the lab-specific self-efficacy, belonging, and recognition items, though we can compare the shifts in PLIC and perceived agency scores to previous studies using these same constructs.

The raw gains in PLIC scores ranged from 0.01 to 0.07 (Table VI), compared with the average raw gains of 0.08 across various types of labs reported in Ref. [19].⁴ The presurvey scores in the current study were on par with those reported in Ref. [19], but the postsurvey scores were lower, suggesting instruction in these courses may have

been on the lower end of effectiveness at shifting students' critical thinking scores.

The raw gains in the perceived agency items ranged from 0.02 to 0.09 (out of 1; Table VI), compared with the average raw gains of 0.2 reported in Ref. [21]. The presurvey perceived agency scores in our study were higher (around 0.7) than those reported in Ref. [21] (which were around 0.5). Our postsurvey scores were similar to theirs (around 0.75 and 0.70, respectively). This suggests that instruction in both studies reached similar levels of perceived agency, or that the survey items reached a ceiling.

We do not see large differences in score increases between courses, which can likely be attributed to the courses using a similar pedagogy [16]. We see several limitations in the interpretability of these results, however. First, the self-efficacy and perceived agency scores show clear ceiling effects. These measures may need to be recalibrated to have a more normal distribution centered on 0.5. Future work should evaluate these measures in other pedagogical and instructional contexts to evaluate the generalizability of this ceiling effect as well as to evaluate the range of student gains across instructional lab pedagogies.

C. Group role effects

Our main result is the impact of the group work role distributions on student outcomes. First, we find no effect of role distribution on PLIC outcomes. Second, we find a consistent trend (with varied significance) across role distributions for the attitude constructs: sharing roles has the strongest positive effect, splitting has the strongest negative effect, and the effect of rotating lies in between.

For PLIC scores, based on the ICAP framework, we expected sharing roles to have a more positive effect than rotating, interpreting sharing roles as better for fostering constructive dialogue than other role distributions. We also expected splitting roles to have the smallest impact, as the division of roles would support individual experts and not necessitate constructive dialogue. Compared to splitting roles, rotating roles means that all students develop expertise across roles, creating opportunities for constructive dialogue between students. Our results, however, show no significant differences in the effects of role distribution on PLIC scores.

This null effect suggests that group work role distributions do not differentially affect the development of critical thinking as measured by the PLIC. We consider two potential explanations. The first relates to measurement precision. The gains in PLIC scores of this sample of students may be too small to allow us to detect effects of role distribution; the mean PLIC scores only increase from 0.01 to 0.07 (out of 1) across courses. The second potential explanation is that groups' *hands-on* role distributions do not affect critical thinking. While students may have distributed the hands-on aspects of roles one way, they might have distributed the *decision-making* aspects differently. For example, a group of students may have split into

⁴Note: Two items were removed from the PLIC after Ref. [19] was published based on work in Ref. [20]. The average raw gain of 0.08 was estimated from the reported item scores, excluding the two removed items.

an equipment handler, a note-taker, and a data analyst for hands-on roles, but may have jointly decided how they would take measurements, write their results, and present their data. In this case, students would have limited hands-on access to roles, but still gain experience in the critical thinking associated with each role. In the ICAP framework, students collaboratively making decisions regarding different roles in the lab would be classified as interactive, the tier with the highest engagement. Thus, it is possible that the distribution of decision making for roles has a greater impact on critical thinking than the distribution of hands-on roles, a distinction that would not be captured by our survey item.

For self-efficacy, we find that sharing roles had a positive, statistically significant effect and splitting roles had a negative, statistically significant effect. Rotating roles had a positive—but not statistically significant—effect. We expect that sharing roles had a net positive effect because each student spends their time on each role every lab session. When rotating roles, a student spends their time on one role each lab session. When splitting roles, students spend their time on only one role throughout the semester. We expect that students who spend more time in a role would have greater self-efficacy in that role. Many of the self-efficacy items in the survey ask about student confidence regarding specific tasks; for example, students are asked if they can “comfortably take charge of the equipment” and “interpret graphs.” On average, students in groups who share more frequently should feel confident with a larger variety of tasks. Conversely, on average, students in groups who only gain experience with one role may feel less confident with tasks outside their role.

The trends in perceived agency are similar to those seen with self-efficacy; though none of the effects are statistically significant for perceived agency. These similar trends match our expectations because the two attitude constructs are theoretically connected [47].

The shift in significance may have several explanations. First, most of the self-efficacy items probe students about specific tasks, while the perceived agency items are more broad. Plausibly, students are answering these items considering their perceived agency in the course overall, which would be similar across all role distributions. Any differences from the group work role distributions would need to be measured by role-specific items so that students evaluate their perceived agency within their lab group. Future work could test the theory that the specificity of self-efficacy items leads to more pronounced gains compared to perceived agency by measuring the gain in each self-efficacy item against the amount of time a student spends in each role or by crafting role-specific perceived agency items that reflect the self-efficacy items.

For belonging, none of the group work role distributions had a statistically significant effect on belonging, though

we again see that sharing had a positive effect and splitting had a negative effect. Research in cooperative learning [10] suggests that group processing [48], through targeted reflection activities, impacts students’ sense of belonging [49]. One possible explanation is that sharing roles, or being in a group that shares roles, is more likely to create opportunities for group processing than the other group work role distributions, explaining the more positive impact on belonging. We may not have observed a statistically significant effect because the labs did not implement a structure to ensure students cooperated and engaged in explicit group processing.

For recognition, the statistically significant, negative effect of splitting on recognition could be explained by students who are siloed into specific roles not feeling like they are doing a “physics” role. This siloing may lead some students to feel that their peers and instructors do not see them as physics people. We would then expect recognition gains to vary based on a student’s role in their groups when they split roles. Based on the findings in Refs. [50–53], we expect that roles like note-taking would have lower recognition gains than roles like equipment handling. Future work could test this theory by measuring recognition gains against the amount of time a student spends in each role. Alternatively, aspects of recognition may be associated with feelings of acceptance by their peers, which is connected to feelings of belonging in cooperative learning [49]. This explanation supports our findings above that the patterns for recognition are similar to those for belonging.

Altogether, these findings and their associated limitations motivate a range of future research. First, because our results are observational, our findings are inevitably intertwined with student choice and preference. Future studies should implement random assignment of lab groups to share, rotate, or split their lab roles throughout a semester to disentangle the isolated effect of group choice on outcome. Research should also expand to more lab environments with different group sizes, lab pedagogies, and student populations. Future work could also use observations of student work in lab instead of student self-reports. Importantly, these results support the findings of Ref. [6] with a larger dataset and additional measurement constructs.

ACKNOWLEDGMENTS

We would like to thank Andrew Loveridge and Viranga Perera for assisting with data collection by distributing surveys. We would also like to thank Matt Thomas and the Cornell Statistical Consulting Unit for their assistance with data analysis. We are very grateful to many discussions with Emily Stump regarding the development of the roles distribution survey item. Finally, we would like to thank Z. Yasemin Kalender and Martin Stein for their work on survey development. This material is based upon work supported by the National Science Foundation Grants No. DGE-2139899 and No. DUE-1836617, the Cornell

Active Learning Initiative, and the Cornell Nexus Scholars Program.

DATA AVAILABILITY

The data that support the findings of this article are openly available [33].

APPENDIX: ASSUMPTIONS FOR HIERARCHICAL LINEAR MODELING

Here, we check the assumptions required for hierarchical linear modeling following prior recommendations [43]. In the main text, we analyze 15 different regression models: three role distributions for five outcome constructs. Here, we provide plots for only the three role distributions for a single outcome construct: the PLIC. The other figures are available in Ref. [33].

We analyze regression models for Eq. (1) in the case of sharing roles, as in Eq. (A1).

$$\text{Post-PLIC}_{ij} = \beta_{0j} + \beta_{1j}\text{Pre-PLIC} + \beta_{2j}\text{Share} + r_{ij}. \quad (\text{A1})$$

Next, in the case of rotating roles, as in Eq. (A2),

$$\text{Post-PLIC}_{ij} = \beta_{0j} + \beta_{1j}\text{Pre-PLIC} + \beta_{2j}\text{Rotate} + r_{ij} \quad (\text{A2})$$

And finally, in the case of splitting roles, as in Eq. (A3).

$$\text{Post-PLIC}_{ij} = \beta_{0j} + \beta_{1j}\text{Pre-PLIC} + \beta_{2j}\text{Split} + r_{ij} \quad (\text{A3})$$

We do not know of any recommendations regarding whether to check the assumptions of hierarchical linear modeling using data before or after multiple imputation. Following another study [54], we check our assumptions using the averaged results from the pooled dataset.

We check the assumption of linearity in Fig. 6. We see no trends in these data, suggesting it meets the assumption of normality. For the other constructs, we do find some ceiling and floor effects (as seen in the original data in Fig. 4) but do not see any other trends in the plots.

For the assumption of homogeneity of variance, we performed analysis of variance and found no statistically significant results (share returned $p = 0.7399$, rotate

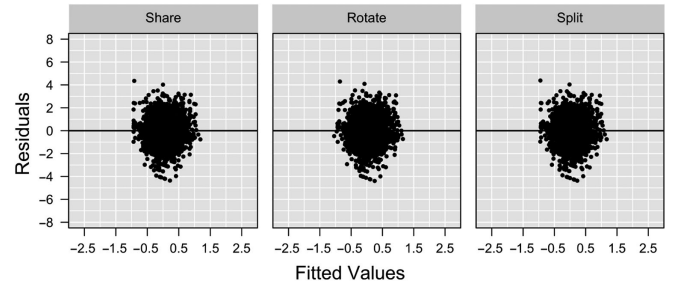


FIG. 6. A visual check for hierarchical linear modeling's assumption of linearity for the PLIC models using data after multiple imputation. These plots display the residuals versus fitted values for three different models: share [see Eq. (A1)], rotate [see Eq. (A2)], and split [see Eq. (A3)].

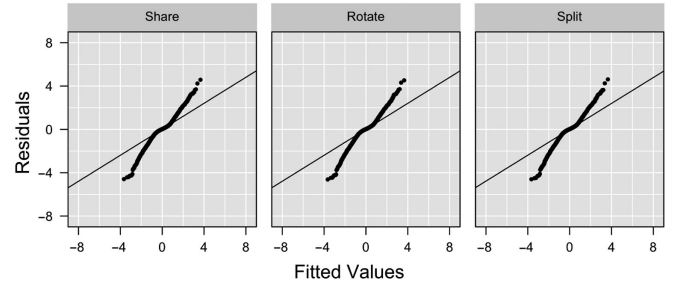


FIG. 7. A visual check for hierarchical linear modeling's assumption of normality for the the PLIC models using data after multiple imputation. These plots display the residuals versus fitted values for three different models: share [see Eq. (A1)], rotate [see Eq. (A2)], and split [see Eq. (A3)].

returned $p = 0.7399$, and split returned $p = 0.7387$). We found no statistically significant effects for any of the three role distribution models across the four attitude constructs ($p \geq 0.5674$), indicating that our data meet the assumption of homogeneity of variance in all cases.

We checked the assumption of normality in Fig. 7. While we see the residuals stray from normality at the ends of the distribution, this usually does not affect interpretations of significance [55,56] and has been seen in similar physics education research studies [4,5,56]. We see somewhat similar trends for our attitude constructs.

We used the `lmer` function from the LME4 package in R for our hierarchical linear models [57].

[1] G. Geschwind, M. Alemani, M. F. Fox, P. Logman, E. Tufino, and H. Lewandowski, Development of a global landscape of undergraduate physics laboratory courses, *Phys. Rev. Phys. Educ. Res.* **20**, 020117 (2024).

[2] 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).

- [3] 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).
- [4] M. Dew, E. Hunt, V. Perera, J. Perry, G. Ponti, and A. Loveridge, Group dynamics in inquiry-based labs: Gender inequities and the efficacy of partner agreements, *Phys. Rev. Phys. Educ. Res.* **20**, 010121 (2024).
- [5] M. Dew, E. M. Stump, and N. Holmes, Structuring groups for gender equitable equipment usage in labs, *Phys. Rev. Phys. Educ. Res.* **21**, 010162 (2025).
- [6] D. Doucette and C. Singh, Share it, don't split it: Can equitable group work improve student outcomes?, *Phys. Teach.* **60**, 166 (2022).
- [7] S. V. Rosser, Group work in science, engineering, and mathematics: Consequences of ignoring gender and race, *Coll. Teach.* **46**, 82 (1998).
- [8] P. Heller and M. Hollabaugh, Teaching problem solving through cooperative grouping. Part 2: Designing problems and structuring groups, *Am. J. Phys.* **60**, 637 (1992).
- [9] Y. Chang and P. Brickman, When group work doesn't work: Insights from students, *CBE Life Sci. Educ.* **17**, ar42 (2018).
- [10] D. W. Johnson and R. T. Johnson, An educational psychology success story: Social interdependence theory and cooperative learning, *Educ. Res.* **38**, 365 (2009).
- [11] D. W. Johnson and R. T. Johnson, Making cooperative learning work, *Theory Pract.* **38**, 67 (1999).
- [12] D. W. Johnson, R. T. Johnson, and K. A. Smith, Cooperative learning returns to college: What evidence is there that it works?, *Change* **30**, 26 (1998).
- [13] K. Tanner, L. S. Chatman, and D. Allen, Approaches to cell biology teaching: Cooperative learning in the science classroom—beyond students working in groups, *Cell Biol. Educ.* **2**, 1 (2003).
- [14] K. D. Tanner, Structure matters: Twenty-one teaching strategies to promote student engagement and cultivate classroom equity, *Cell Biol. Educ.* **12**, 322 (2013).
- [15] M. T. H. Chi and R. Wylie, The ICAP framework: Linking cognitive engagement to active learning outcomes, *Educ. Psychol.* **49**, 219 (2014).
- [16] E. M. Smith, M. M. Stein, C. Walsh, and N. G. Holmes, Direct measurement of the impact of teaching experimentation in physics labs, *Phys. Rev. X* **10**, 011029 (2020).
- [17] L. B. Buck, S. L. Bretz, and M. H. Towns, Characterizing the level of inquiry in the undergraduate laboratory, *J. Coll. Sci. Teach.* **38**, 52 (2008).
- [18] N. Holmes and E. M. Smith, Operationalizing the AAPT learning goals for the lab, *Phys. Teach.* **57**, 296 (2019).
- [19] C. Walsh, K. N. Quinn, C. Wieman, and N. G. Holmes, Quantifying critical thinking: Development and validation of the physics lab inventory of critical thinking, *Phys. Rev. Phys. Educ. Res.* **15**, 010135 (2019).
- [20] A. B. Heim, C. Walsh, D. Esparza, M. K. Smith, and N. G. Holmes, What influences students' abilities to critically evaluate scientific investigations?, *PLoS One* **17**, e0273337 (2022).
- [21] Z. Y. Kalender, E. Stump, K. Hubenig, and N. G. Holmes, Restructuring physics labs to cultivate sense of student agency, *Phys. Rev. Phys. Educ. Res.* **17**, 020128 (2021).
- [22] Z. Y. Kalender, M. M. Stein, and N. Holmes, Sense of agency, gender, and students' perception in open-ended physics labs, *presented at PER Conf. 2020, Virtual Conference*, 10.1119/perc.2020.pr.Kalender.
- [23] L. E. Kost-Smith, Characterizing, modeling, and addressing gender disparities in introductory college physics, Ph.D. thesis, University of Colorado, Boulder, 2011.
- [24] Z. Y. Kalender, E. Marshman, C. D. Schunn, T. J. Nokes-Malach, and C. Singh, Gendered patterns in the construction of physics identity from motivational factors, *Phys. Rev. Phys. Educ. Res.* **15**, 020119 (2019).
- [25] A. Tapal, E. Oren, R. Dar, and B. Eitam, The sense of agency scale: A measure of consciously perceived control over one's mind, body, and the immediate environment, *Front. Psychol.* **8** (2017).
- [26] M. Scardamalia and C. Bereiter, Knowledge building and knowledge creation: Theory, pedagogy, and technology, in *The Cambridge Handbook of the Learning Sciences*, edited by R. K. Sawyer and C. Bereiter (Cambridge University Press, Cambridge, England, 2014), pp. 397–417.
- [27] C. Good, A. Rattan, and C. S. Dweck, Why do women opt out? Sense of belonging and women's representation in mathematics, *J. Pers. Soc. Psychol.* **102**, 700 (2012).
- [28] R. M. Lock, Z. Hazari, and G. Potvin, Impact of out-of-class science and engineering activities on physics identity and career intentions, *Phys. Rev. Phys. Educ. Res.* **15**, 020137 (2019).
- [29] W. K. Adams and C. E. Wieman, Development and validation of instruments to measure learning of expert-like thinking, *Int. J. Sci. Educ.* **33**, 1289 (2011).
- [30] E. Knehta, C. Runyon, and S. Eddy, One size doesn't fit all: Using factor analysis to gather validity evidence when using surveys in your research, *CBE Life Sci. Educ.* **18**, 1 (2019).
- [31] T. Raykov and G. A. Marcoulides, *An Introduction to Applied Multivariate Analysis* (Routledge, New York, NY, 2008).
- [32] B. G. Tabachnick, L. S. Fidell, and J. B. Ullman, *Using Multivariate Statistics* (Pearson, Boston, MA, 2007), Vol. 5.
- [33] J. Feinleib, M. Dew, and N. G. Holmes, Share, rotate, split: The effects of group work role distributions on student outcomes (2026), 10.17605/OSF.IO/8SKHU.
- [34] R. B. Kline, *Principles and Practice of Structural Equation Modeling* (Guilford Publications, New York, NY, 2023).
- [35] J. Nissen, R. Donatello, and B. Van Dusen, Missing data and bias in physics education research: A case for using multiple imputation, *Phys. Rev. Phys. Educ. Res.* **15**, 020106 (2019).
- [36] A. D. Woods, D. Gerasimova, B. Van Dusen, J. Nissen, S. Bainter, A. Uzdavines, P. E. Davis-Kean, M. Halvorson, K. M. King, J. A. R. Logan, M. Xu, M. R. Vasilev, J. M. Clay, D. Moreau, K. Joyal-Desmarais, R. A. Cruz, D. M. Y. Brown, K. Schmidt, and M. M. Elsherif, Best practices for addressing missing data through multiple imputation, *Infant Child Dev.* **33**, e2407 (2024).
- [37] A. D. Woods, P. Davis-Kean, M. A. Halvorson, K. M. King, J. A. R. Logan, M. Xu, S. Bainter, D. M. Y. Brown, J. M. Clay, R. A. Cruz *et al.*, Missing data and

- multiple imputation decision tree, PsyArXiv (2021), [10.31234/OSF.IO/MDW5R](https://doi.org/10.31234/OSF.IO/MDW5R).
- [38] J. W. Graham and J. L. Schafer, On the performance of multiple imputation for multivariate data with small sample size, in *Statistical Strategies for Small Sample Research* (Sage, Thousand Oaks, CA, 1999), Vol. 50, p. 1.
- [39] S. van Buuren and K. Groothuis-Oudshoorn, MICE: Multivariate imputation by chained equations in R, *J. Stat. Softw.* **45**, 1 (2011).
- [40] J. R. van Ginkel, M. Linting, R. C. A. Rippe, and A. van der Voort, Rebutting existing misconceptions about multiple imputation as a method for handling missing data, *J. Pers. Assess.* **102**, 297 (2020).
- [41] P. Madley-Dowd, R. Hughes, K. Tilling, and J. Heron, The proportion of missing data should not be used to guide decisions on multiple imputation, *J. Clin. Epidemiol.* **110**, 63 (2019).
- [42] I. R. White, P. Royston, and A. M. Wood, Multiple imputation using chained equations: Issues and guidance for practice, *Stat. Med.* **30**, 377 (2011).
- [43] 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).
- [44] E. Theobald, Students are rarely independent: When, why, and how to use random effects in discipline-based education research, *CBE Life Sci. Educ.* **17**, rm2 (2018).
- [45] M. Dew, A. M. 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](https://doi.org/10.1119/perc.2022.pr.Dew).
- [46] A. L. Traxler, X. C. Cid, J. Blue, and R. Barthelemy, Enriching gender in physics education research: A binary past and a complex future, *Phys. Rev. Phys. Educ. Res.* **12**, 020114 (2016).
- [47] A. Bandura, Self-efficacy mechanism in human agency, *Am. Psychol.* **37**, 122 (1982).
- [48] A. Bertucci, S. Conte, D. W. Johnson, and R. T. Johnson, Influence of group processing on achievement and perception of social and academic support in elementary inexperienced cooperative learning groups, *J. Educ. Res.* **105**, 329 (2012).
- [49] M. F. Strahm, Cooperative learning: Group processing and students needs for self-worth and belonging, *Alberta J. Educ. Res.* **53**, 63 (2007).
- [50] 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).
- [51] H. Pettersson, Making masculinity in plasma physics: Machines, labour and experiments, *Sci. Technol. Stud.* **24**, 47 (2011).
- [52] A. T. Danielsson, Exploring woman university physics students "doing gender" and "doing physics", *Gender Educ.* **24**, 25 (2012).
- [53] A. J. Gonsalves, A. T. Danielsson, and H. Pattersson, Masculinities and experimental practices in physics: The view from three case studies, *Phys. Rev. Phys. Educ. Res.* **12**, 020120 (2016).
- [54] J. M. Nissen, I. Her Many Horses, and B. Van Dusen, Investigating society's educational debts due to racism and sexism in student attitudes about physics using quantitative critical race theory, *Phys. Rev. Phys. Educ. Res.* **17**, 010116 (2021).
- [55] 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).
- [56] C. Walsh, H. J. Lewandowski, and N. G. Holmes, Skills-focused lab instruction improves critical thinking skills and experimentation views for all students, *Phys. Rev. Phys. Educ. Res.* **18**, 010128 (2022).
- [57] D. Bates, M. Mächler, B. Bolker, and S. Walker, Fitting linear mixed-effects models using LME4, *J. Stat. Softw.* **67**, 1 (2015).