Editors' Suggestion

Group roles in unstructured labs show inequitable gender divide

Katherine N. Quinn,[1,2] Michelle M. Kelley,[3] Kathryn L. McGill,[4] Emily M. Smith,[5] Zachary Whipps,[6] and N. G. Holmes[3]

[1]*Center for the Physics of Biological Function, Princeton University, Princeton, New Jersey 08540, USA*
[2]*Initiative for the Theoretical Sciences, CUNY Graduate Center, New York, New York 10016, USA*
[3]*Laboratory of Atomic and Solid State Physics, Department of Physics, Cornell University, Ithaca, New York 14853, USA*
[4]*Department of Physics, University of Florida, Gainesville, Florida 32611, USA*
[5]*Department of Physics, Colorado School of Mines, Golden, Colorado 80401, USA*
[6]*Department of Physics, Cornell University, Ithaca, New York 14853, USA*

(Received 20 December 2019; accepted 28 April 2020; published 26 May 2020)

Instructional labs are being transformed to better reflect authentic scientific practice, often by removing aspects of pedagogical structure to support student agency and decision making. We explored how these changes impact men's and women's participation in group work associated with labs through clustering methods on the quantified behavior of students. We compared the group roles students take on in two different types of instructional settings: (i) highly structured traditional labs, and (ii) less structured inquiry-based labs. Students working in groups in the inquiry-based (less structured) labs assumed different roles within their groups, however men and women systematically took on different roles and men behaved differently when in single- versus mixed-gender groups. We found no such systematic differences in role division among male and female students in the traditional (highly structured) labs. Students in the inquiry-based labs were not overtly assigned these roles, indicating that the inequitable division of roles was not a result of explicit assignment. Our results highlight the importance of structuring equitable group dynamics in educational settings, as a gendered division of roles can emerge without active intervention. As the culture in physics evolves to remove systematic gender biases in the field, instructors in educational settings must not only remove explicitly biased aspects of curricula but also take active steps to ensure that potentially discriminatory aspects are not inadvertently reinforced.

DOI: 10.1103/PhysRevPhysEducRes.16.010129

I. INTRODUCTION

The demographic composition of physicists is not representative of the general population, with men overrepresented not only in number but also in high-ranking positions within the physics community [1]. In exploring the underlying mechanisms for this, there has been a large focus in education research on gaps in performance between men and women on concept inventories and course grades [2,3]. While informative, this approach provides an incomplete picture [3,4]; importantly, student persistence in physics can often be independent of their physics test scores [5]. New strides in science education research now include investigating more metrics such as sociocultural factors [6,7], self-efficacy [8,9], sense of belonging [10], and identity formation [9,11,12]. Moreover, participation in the physics community through the *roles* people take on within the community can heavily shape one's identity as a physicist [13]. Within any community, members assume different roles as they take on different responsibilities, perform certain functions, and are perceived in specific ways by themselves and by the group [14]. Understanding what roles develop throughout students' physics education is critical, as the field of physics is associated with masculinity, suggesting that a gendered division of roles may greatly influence the modern practice of physics. [15,16].

Students have little direct experience with the field, however [17], and their perceptions of the field and their physics identities are developed through their immersion in physics courses. Many courses (including labs) involve significant group work, which can leverage the fact that strong peer relationships can benefit students' development of their science identities [18–21]. As with group work in other aspects of physics courses (such as cooperative problem solving, tutorial, or in-class lecture activities), lab activities require coordination of group members as they collect and interpret a common dataset. Lab activities are distinct from other learning environments in that there

are multiple distinct activities that must be carried out, so division of labor, and thus assigning distinct roles, is much more common.

In this study, we explored patterns in the behaviors students exhibit in the context of physics labs. In doing so, we aim to better understand the group roles that emerge in these spaces. Labs provide an environment where students interact with peers and engage in physics experiments in ways that can influence their perception of physics and of themselves as physicists [22]. Furthermore, labs are changing nationally in response to calls to provide students with more authentic science experiences [23]. Understanding how students behave and interact with each other in different lab environments, and the roles students assume in these settings, can inform educators and researchers when designing new pedagogy to better address inequities.

Identity formation is a complicated, multidimensional process that includes gender, race, physical ability, socioeconomic status, sexual orientation, and religion, among many, many other factors. The formative process includes individual agency as well as broader cultural and societal factors [24]: the impact of the broader culture outside of the physics classroom strongly influences one's identity formation (such as a culturally perceived notion of physics as a masculine field [25]). Importantly, how one develops a sense of identity impacts the set of available roles one may take on in a particular context, and strongly determines persistence in a particular field [26]. In this study, we analyzed the quantified behavior profiles (discussed further in Sec. II) as a way of probing the roles students take on in physics labs to understand some of the ways in which these roles can be equitably or inequitably divided.

We define an equitable division of roles as one in which all members are equally likely to assume each role, i.e., every role is available to every member. Note that this is different from equal or identical roles, in which every student performs the same function and thus would behave similarly from each other. An inequitable division of roles is one in which not every role is available to every member. For example, certain members are expected to assume, or prevented from taking on, certain roles. At the individual level, roles that are divided among group members may not be indicative of inequitable division. However, if students systematically behave differently in groups, then broader statistical analyses will reveal these overarching inequities. For instance, roles may be *gendered*, in the sense that there is an inequitable gender divide, with men and women taking on systematically different roles [27,28].

Prior research has found that group work often involves inequitable participation between men and women. For example, female students participated less in group discussion when they were outnumbered by male students [29,30] and responded disproportionately less than male students to instructor-posed questions in lecture [31]. In physics lab courses, students have described the available roles themselves as being either masculine or feminine [22] and women have been found to engage less frequently with hands-on equipment [32,33] or with computers [34] when working in mixed-gender pairs. In contrast, contradicting results were found when comparing the performance of male majority, female majority, and mixed groups on engineering design tasks across two different courses [35].

Individual students' behaviors can be used to probe the roles they take on in physics labs, and are likely a result of their personal identity (gender or otherwise), the particular instructional context, and the broader physics culture [26,27,34,36–38]. Understanding students' experiences in labs through the behaviors and roles they take on both highlights existing gender disparities as well as informs future research on students' persistence in science. We specifically sought to understand the impact of *different* instructional lab environments on student roles and how these roles are divided between men and women. One way to make labs more authentic is to make them discovery-based and inquiry-driven, removing structure from the lab. How does removing pedagogical structure in the lab impact these learning environments? Specifically, what impact does pedagogical structure have on the equitable division of roles within groups?

II. MATERIALS AND METHODS

All participants in this study were undergraduate students at a major research university enrolled in the honors-level mechanics course of a calculus-based physics sequence. The course was designed for physics majors and open to students across the sciences and engineering. The sample of prospective physics majors is an important population for this study, given the potential link between students' experiences, roles, identity, and persistence in physics [26]. We explored students' behaviors in two different types of lab instruction.

The highly structured *traditional labs* were designed to reinforce physics content knowledge presented in lecture. Students were provided with detailed paper worksheets to follow during lab, guiding them through experiments that provided them with hands-on experience. The lab guides provided explicit details about what and how much data to collect and posed targeted conceptual physics questions to support making predictions and interpreting results. Students worked in groups to collect data for the experiments and submitted individual paper worksheets.

In contrast, the less structured *inquiry labs* were designed to emphasize the process of experimentation in physics (see, for example, Refs. [39–42]). Students were provided with a specific goal, but were expected to design their own experiment to achieve that goal. Lab guides prompted students to design data collection methods to reflect on results, and to design follow-up investigations to improve or extend their investigations. Students worked collaboratively to design and implement their experiments

and submitted only one electronic notebook as a group. Reference [40] includes additional detail about the differences between the conditions, including differences between students' learning engagement with experimentation, and attitudes towards experimental physics.

The same mechanics course was taught twice during the academic year, once in the fall semester and then again in the spring semester. Students from both semesters were included in this study. During the first semester, all students attended the same lecture, mixed together in discussion sections, but were separated into two pedagogically different lab types discussed below (three traditional lab sections and two inquiry lab sections). During the second semester, the two lab sections under study were both inquiry labs. Note that we observed students across multiple lab periods throughout the course of the semester (and each student appeared in only one semester), and so while each student is in one lab section they appear in multiple lab periods. All participants were unaware of the differences between lab types: students in the first semester self-selected into their lab sections prior to the start of the course by registering for the course, and only the inquiry lab sections were available to students in the second semester. Student groups varied every period, and were randomly assigned.

The role a student takes on in their group is a highly complex reflection of the function they serve in the group, and depends on numerous factors from the individual to the cultural level. Because this study explores student roles in physics labs, we assume that these roles are in some way correlated with their behavior in these labs, such as handling of equipment or of computer usage. To probe the roles that students assumed in physics labs, we analyzed the quantified behavior profiles of 143 students across multiple lab periods. We collected data for this study at two levels of granularity.

First, *coarse behaviors* were captured at five minute intervals for all students in multiple lab periods. The codes were determined by what the students were handling: (1) lab desktop computer, (2) personal laptop or other device, (3) writing on paper, (4) handling equipment, or (5) engaging in some other activity. We used the "other" code to capture all other behaviors, such as discussing within their group, with another group, or with the instructor; engaging in whole-class discussions; writing on whiteboards; or engaging in off-task behaviors. Note that the other code was constructed to ensure all time was coded for every student, and therefore captures many different behaviors. The choice of codes were designed to capture enough detailed information as possible about every student, coded in real time, while reflecting the lack of *a priori* knowledge of what the exact group roles were. The behaviors of each student in each lab period were amassed to create a *profile* of their behaviors during that lab period. Unfortunately, given the observation protocol (discussed in detail in Sec. II B) where each student was observed over the course of an entire lab period, subdividing the other code could not be done quickly enough and with enough accuracy by the researchers. Instead, a second analysis of such detailed behavior was performed using video from single groups, and discussed in greater detail in Sec. II D.

A. Collecting demographic information

We used in-class surveys to obtain student demographic information. In all, 143 students across multiple lab sections were used in this study. While they had the option to disclose a gender other than *woman* or *man*, no student chose to do so, and only two students did not disclose their gender identity. As a result, all students were included in the initial cluster analysis, however the gender analysis follows the traditional gender binary of *woman* or *man* (with the two undisclosed students omitted from the graphs in Figs. 4 and 6 due to insufficient statistics). Table I shows the demographic breakdown of student participants in this study. To obtain the standard error on the fraction of a population (such as in Table I or Fig. 6), we used the following:

$$\mathtt{Err}(p,N)=\sqrt{\frac{p(1-p)}{N}}, \tag{1}$$

where p is the fraction of the population, and N is the size of the total population.

B. Quantifying coarse student behaviors

In all lab sections, observers documented student behaviors following the observation protocol used in Ref. [34]. Every 5 min, an observer noted each student's actions in the lab using one of five codes: desktop, equipment, laptop, paper, and other. One code was applied to each student in the class at each 5 min interval, except in cases where students could not be observed (e.g. because they were late or left early). The codes are described in Table II, and were based on what a student could be handling in the lab. The other code captured all other behaviors such as engaging in whole-class discussions, writing on whiteboards, discussing with the teaching assistant (TA) or undergraduate teaching assistant (UTA), and off-task behaviors, ensuring that all in-lab time was coded. The desktop code was

TABLE I. Student demographics of this study. Errors were computed using standard error for population fractions, shown in Eq. (1). In all, 143 students were considered in this study.

	Traditional labs		Inquiry labs	
	N	%	N	%
Women	11	19 ± 5	21	25 ± 5
Men	46	79 ± 5	63	74 ± 5
Undisclosed	1	2 ± 2	1	1 ± 1

TABLE II. Action codes used in observations. The laptop code is used for both handling a laptop or personal device (students used laptops, phones, and tablets for the purpose of notetaking, write up, data analysis, and reading instructions in the inquiry labs).

Code	Description
Desktop	Using the desktop computer at the lab bench.
Equipment	Handling equipment.
Laptop	Using a laptop or personal device.
Paper	Writing on paper or in a notebook.
Other	Other action or behavior.

separated from the laptop code because the desktop was often required for data collection (e.g., because it was directly connected to a detector or piece of equipment). Furthermore, desktops were shared within groups whereas the laptop code was ascribed to students handling personal devices. While desktops were present in both lab types, only students in the inquiry labs actively used laptops to analyze data, document their lab procedures, and submit their electronic notebooks.

The codes used in this study, in particular the other code, are very coarse and so multiple behaviors can fall under the same code (e.g., the laptop code includes using a laptop for data analysis as well as for note taking, the other code captures activities such as discussing the lab with group members or engaging in off-task talking with group members). Given the observation protocol, it was not possible for an observer to differentiate between these different and more nuanced behaviors in real time for every student, and so a second analysis was performed, and the details of which are outlined in Sec. II D.

To validate our observation procedure, two observers coded student actions in the same lab period using the described protocol but at different 5 min intervals. If we had each observer code the same student at the same time, we would have only evaluated the reliability of the codes. Instead, observers were specifically not coding the same student at the same time. Thus, comparing the overall code count for each student provides a measure of reliability of the codes recorded at 5 min intervals. By comparing the overall code count for each student, we provide a measure of reliability about the overall method. This method limits us, however, from comparing individual student behavior over time in the lab period. Thus, all analysis is performed on the student profiles, which aggregate their behaviors throughout the lab period. Note that because observers were explicitly not observing the same student at the same time, percent agreement or calculating Cohen's kappa would not provide the necessary information to validate the method. Instead, a standard chi-squared analysis was performed on the contingency table constructed from the accumulated codes (the frequency each observer noted each code, summed over all students). We used the criteria that if two sets of observations are statistically indistinguishable from each other, then the observers captured the same overall profiles for the students in the lab session. Note that, if either (i) there was not agreement between the codes, or (ii) the 5 min interval did not accurately capture student behavior when averaged over a lab period, then there would be disagreement in these overall distributions.

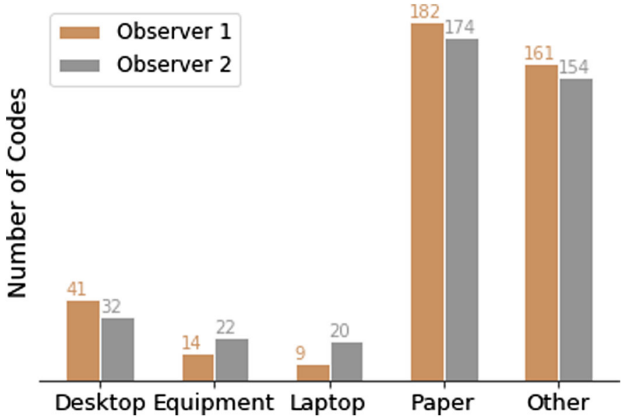

FIG. 1. Bar plot of code counts from two observers used to form the basis of a chi-squared test to validate the observation protocol used in this study. Two observers documented the same lab period, and the resulting contingency table (given by the raw counts displayed on the graph and shown in Table III) was used to determine statistical validity of the method. Here, the two distributions are statistically indistinguishable indicating that the observers captured the same distribution of student actions.

In all cases observers' distributions were statistically indistinguishable, and so single observers coded subsequent lab periods. When attempts were made at subdividing the codes, for instance, to capture students performing data analysis vs note taking or identifying if group discussions were off task, we were not able to obtain agreement between observers. As such, we used the protocol detailed in this section. We provide an example of observer comparisons for illustrative purposes. A sample graph of the accumulated codes for two observers in a traditional lab section is presented in Fig. 1. The contingency table constructed from these observations is given by Table III. Because the two distributions are statistically indistinguishable, the observers captured the same distribution of student actions.

TABLE III. Sample contingency table used to determine if two distributions are statistically different. Two observers documented the same lab period, and a chi-squared test was performed to determine if the resulting distributions are statistically similar or dissimilar. Here, we obtain $p > 0.1$, indicating that the observers captured the same distribution of student actions.

Observer	Desktop	Equipment	Laptop	Paper	Other
1	41	14	9	182	161
2	32	22	20	174	154

TABLE IV. Demographic breakdown of student profiles measured in this study. Errors were computed using standard error for population fractions, shown in Eq. (1). In all, 143 students were observed across multiple lab periods, resulting in 522 unique student profiles.

	Traditional labs		Inquiry labs	
	N	%	*N*	%
Women	34	18 ± 3	87	26 ± 2
Men	152	81 ± 3	246	74 ± 2
Undisclosed	2	1 ± 1	1	0.3 ± 0.3

Because students were observed during multiple lab periods over a full semester, we were able to document individual students more than once. As a result, we obtained 522 unique student profiles, each quantifying the actions of one student in one lab period through the frequency of associated codes. Table IV shows a demographic breakdown of the student profiles used in this study.

C. Cluster analysis

The distribution of coarse behavior code frequencies are highly skewed, with most students engaging in a particular activity infrequently or not at all and some students engaging in an activity a lot. Figure 2 shows box plots of the raw data, illustrating the non-Gaussian features of the data. For this reason, we performed a cluster analysis instead of methods that rely on the assumption of Gaussian distributions. Clustering can account for nonlinearities

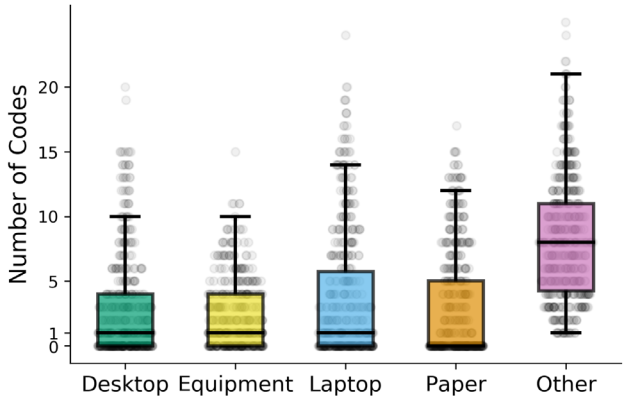

FIG. 2. Box plots of raw data revealing the highly non-Gaussian nature of the code distributions. Each faded point is the accumulated codes for a student in a lab period for a particular category (the horizontal spread of the points is just to visualize all the points), and so darker regions represent more total codes of that value (with the darkest regions near zero). Note that the median for all codes except other is less than or equal to one, reflecting the fact that over half of students were observed engaging in that behavior once or less than once. This, combined with the fact that there are a large number of outliers, is an indication that students either engage in a particular activity a lot or not at all.

missed in common regression analyses, capturing *dominant* behavior as opposed to *average* behavior, and has been used in similar studies of this type to provide fruitful results [43]. By performing a demographic analysis on the student groupings (i.e., clusters) we can quantitatively characterize coarse gendered behavior.

To perform a cluster analysis on multidimensional data, the scales for each measure must be the same. In this study, there were two major effects present that caused differences in scales that we accounted for. First, the amount of coded time for each student was highly variable, ranging from less than 45 min to over 175 min. To account for this effect, we normalized each student profile. In this way, each measure represents the fraction of time spent on a particular task. Second, there is the inherent differences in the five measures. For instance, from Fig. 2, we can see that the distributions for other is more spread out than for equipment. To account for this, each measure was grand mean scaled so that, averaged over all students, each measure had mean 0 and standard deviation of 1. In doing so, each measure becomes a Z score [43,44]. Thus, each student's Z score tells us whether the time they spent on a particular activity was above or below average as compared to other students. Moreover, the Euclidean distance between two profiles has a statistical interpretation in this Z-score format: it measures the dissimilarity of two student profiles in units of standard deviations [43].

We performed a standard k-means clustering on the rescaled student profiles. k-means clustering is an iterative algorithm, where the researcher specifies the number of clusters. The algorithm clusters and then reclusters the data in an iterative manner until the sum of the square of the distances from all points to their respective cluster's center is minimized and no point changes cluster between iterations [45].

Note that not all data can be meaningfully clustered. For example, even if all data form a structureless blob, a researcher can still input two or more clusters and the algorithm will converge to a solution. Therefore, in order to determine (i) if the data are clusterable, and (ii) if so, what the optimal number of clusters is, we used the elbow method [46]. We plotted the average squared distance from each point to the center of its assigned cluster, as a function of the number of clusters, and compared the results to 10 000 randomly generated student profiles. We used enough random data to numerically generate a smooth function and ensure that the comparison is not hindered by statistical fluctuations. The results of the elbow plot are shown in Fig. 3. The plot for our collected data was substantially below random, indicating that the data is clusterable. There is a distinct kink in the plot for five clusters, indicating that the optimal number of clusters is five.

From the elbow plot in Fig. 3, specifically from looking at the drop in average squared distance from each point to the center of its cluster for five clusters compared to one, we

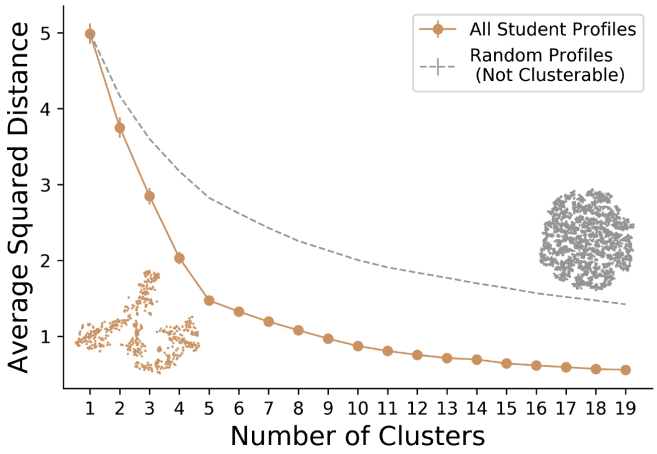

FIG. 3. Elbow plot used to determine the optimal number of clusters for the data. The average squared distance from each point to the center of its assigned cluster is plotted as a function of the number of clusters. There is a kink at five, indicating that the optimal number of clusters for the data is five. Our results were compared against 10 000 randomly generated student profiles. Note that the elbow is well below random, a sign that the data can be clustered. Superimposed on the graph is a two-dimensional visualization of the data and random points for qualitative comparison. The data show structure (brown points in lower left), whereas the random points form a blob (gray points in center right).

can see that the five optimal clusters account for 70% of the variance in the data. By looking at the distances confined to each of the five measures (i.e., generating similar figures as that of Fig. 3 for each measure, where the max value would be 1 instead of 5), we found that the five optimal clusters account for 73% of desktop use, 60% of equipment use, 78% of laptop use, and 59% of other activities). This is well above the 50% threshold used for a study of this type [43,44].

We provide a 2D visualization of the clusters using t-SNE [47], with each dot representing a profile colored by its assigned cluster (Fig. 4). Figure 4 is a two-dimensional representation of a five-dimensional space, and so is used primarily for qualitative illustration.

Clusters from k means are characterized by their centers. Here, the centers of the five clusters matched the five codes used in this study and so we labeled the clusters accordingly. Therefore, the clusters characterize "high users" of a particular measure. Note that this description fits with the raw data, shown in Fig. 2, which illustrates that the majority of students engage in a particular task either frequently or very rarely. For example, students in the yellow cluster of Fig. 4 spent a larger fraction of their time on the equipment than the average student, so this cluster is referred to as the equipment cluster. This is a feature of student behaviors,

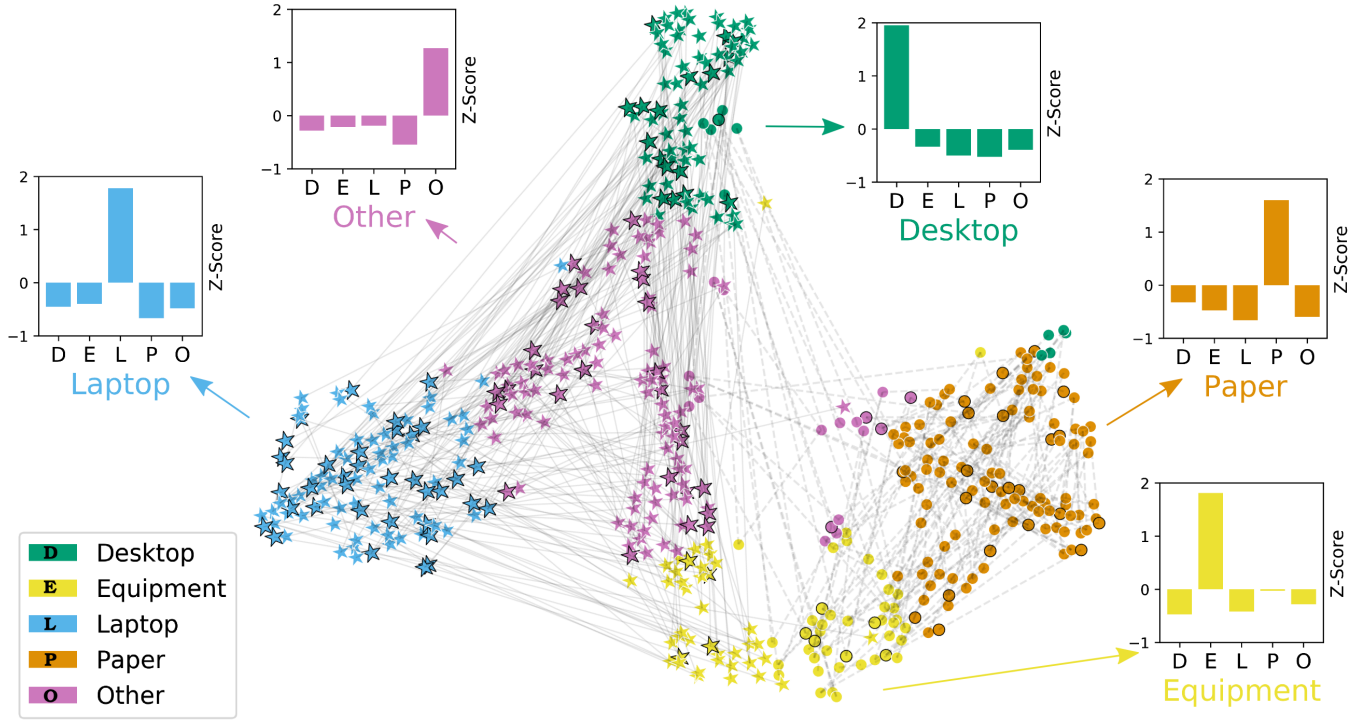

FIG. 4. Two-dimensional visualization of behavior clusters and their centers. Each point represents a unique student profile, with profiles from the same group connected by a gray line (solid for less-structured inquiry labs, and dashed for highly structured traditional labs). Circles represent students in the traditional labs and stars in the inquiry labs, and black edges indicate women's profiles. All points in the laptop cluster are stars, whereas all points in the paper cluster are circles, a reflection of the pedagogical differences in the labs (students in the traditional labs were filling out paper worksheets, whereas in the inquiry labs were filling out electronic notebooks). Clusters are characterized by their centers, and here the centers of the five clusters are given by large Z scores for each of our codes.

and not due to the number of codes used in this study. For instance, one could imagine a scenario in which all students behave nearly identically, with minor differences described by fluctuations: in that case, the data would form a five-dimensional Gaussian cloud centered at zero, and an elbow plot that matches random noise. Or, one could imagine a scenario in which only students handling equipment handle the lab desktop, in which case a cluster would emerge that couples the two respective codes.

We used the clusters that emerged from the data to coarsely characterize the roles students take on in labs. Generally, roles within groups are complex and multidimensional and could be further explored in greater detail through more detailed video analysis (discussed in Sec. II D), student interviews, or anthropological investigations. The analysis performed here provides a coarse-grained perspective on the division of roles within groups, and will ultimately reveal the unexpected inequities in role divisions (discussed next).

Because each student had multiple profiles, arising from several lab periods over the course of a semester, we investigated whether or not it is possible to further collapse the profiles to determine "semester-long" behaviors. We did this by analyzing whether or not individual students' profiles appear in multiple clusters over the course of a semester. In the traditional labs, $87\% \pm 4\%$ of students have profiles appearing in more than one cluster. Similarly, $86\% \pm 4\%$ of students in the inquiry lab appear in more than one cluster. Because so many students have profiles appearing in multiple clusters, the weekly variation in an individual's profile is too great to further collapse (for numerous reasons, such as variability in lab content and students changing lab partners).

D. Describing detailed student behavior

We used video recording of single groups during full lab periods to better describe student behavior in more detail than captured in the previous section. In all, ten videos were coded, decomposing 23 profiles from 17 students (five students appeared in more than one video). BORIS software was used to code videos [48], specifically the fraction of time students engaged in different behaviors.

The five codes in Table II were further broken down by what a student was doing (e.g., analyzing data) while engaged in that coarse behavior (e.g., using the desktop) as shown in Fig. 5. The paper code was used to predominantly describe students filling out paper worksheets in the traditional labs, and so it was not further decomposed. Students in the inquiry labs predominantly used whiteboards for calculations, and very rarely used paper. Both the desktop and laptop codes were used to describe students analyzing data, collecting data, or writing lab notes, and so both of these codes were broken down in this way.

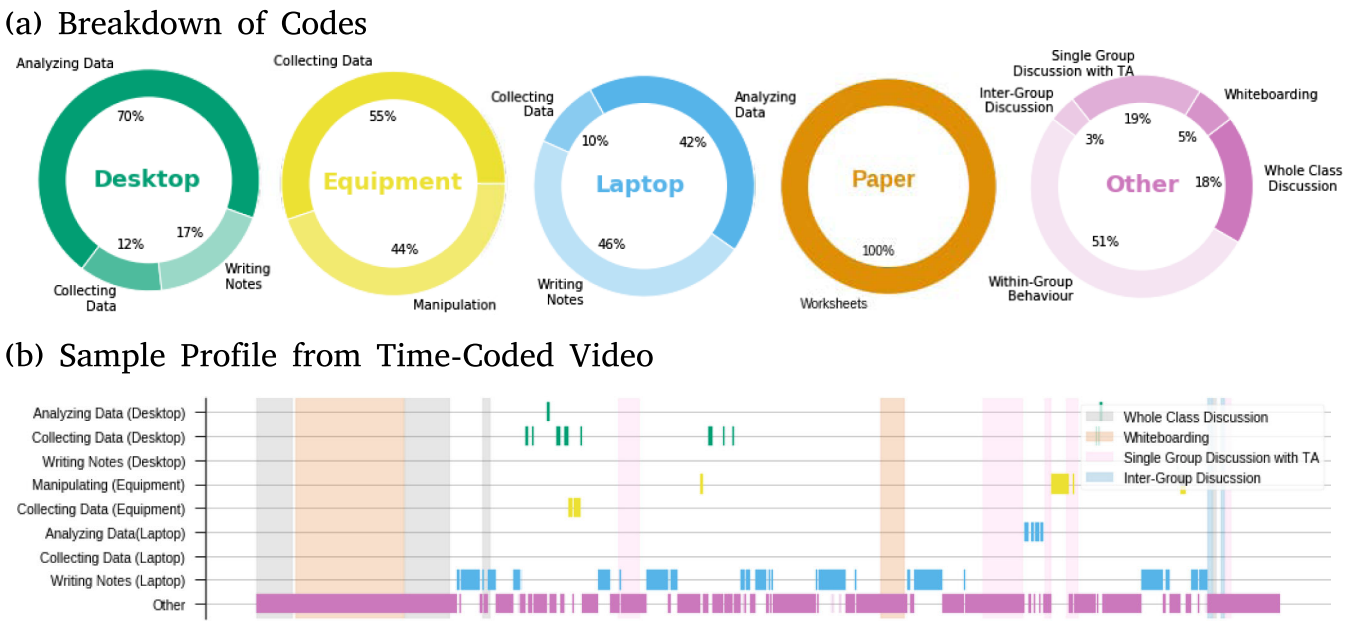

FIG. 5. Breakdown of codes by decomposing coarse behavior (e.g., "handling laptop") into more fine-grained behavior (e.g., "analyzing data"). Ten videos were coded, resulting in 23 decomposed profiles from 17 different students (five students appeared in more than one video). (a) A breakdown of each code, showing the fraction of time students engaged in a particular task while coded as a particular behavior. Three of the five codes (desktop, equipment, and laptop) were directly decomposed into subcodes while analyzing videos, as shown in (b) illustrating a sample coded time series. Four additional "group states" were coded in the videos, representing large group behavior (discussing with a TA or UTA, conversing with other groups, whole class discussions and announcements, and using a whiteboard). We decomposed the other code by overlapping it with these larger group states. The paper code was purely represented by students filling out paper worksheets in the traditional labs.

TABLE V. Event codes used in video observations. These codes described significant events in the lab, and were used to decompose the more coarse-grained other code. A sample time series illustrating a coded video is shown in Fig. 5(b).

Code	Description
Whole class discussion	The TA or UTA makes an announcement to the class, or holds a whole class discussion.
Whiteboarding	Students perform invention activities in the lab, and use a white board to sketch out ideas and concepts.
Single group discussion with the TA	TA or UTA engages in a discussion with the group (but not as part of a whole class discussion).
Intergroup discussion	Groups compare results or discuss among each other (not as part of a whole class discussion).

However, when collecting data, the desktop was often connected directly to equipment whereas gathering data on a laptop was purely represented by students manually entering data into their electronic notebook or analysis software. Students handling equipment were primarily doing so to either collect data or manipulate the setup in some way (setup, cleanup, calibration, playing) and so the equipment code can be further decomposed into these two tasks. In this way, the desktop, equipment, laptop, and paper codes were explicitly decomposed.

To better describe student behavior while coded as other, we introduced four new state codes. These were used to describe significant events in lab, and are elaborated in Table V. By overlapping the event codes with other, we broke down the other code and provide a more qualitative picture of classroom activities, such as engaging in whole-class discussions, using whiteboards to sketch out ideas and concepts, single group discussions with the TA or UTA, or engaging in intergroup discussions with neighboring groups.

To validate this method, two observers coded the same video as a means of testing the interrater reliability. The level of agreement was assessed with Cohen's kappa where a value of 0.61–0.80 represents substantial agreement. Two observers coded the same video, and obtained a Cohen's kappa value of 0.79, indicating substantial agreement between the two. As a result, only one researcher coded the subsequent videos.

Video analysis was also used to better understand task allocation. Point events were identified when one student explicitly instructed another to perform a task. We broke down the criteria for inclusion as a point event and exclusion as a point event in the following way:

- *Criteria for inclusion*: A student needs to be addressing another, and explicitly direct them in some way, such as by saying "You should do X."
- *Criteria for exclusion*: Suggesting a task should be done that a student assumes without being asked is not included. Examples of such events are characterized by statements such as "We should do X," "I think we should focus on X," "Does someone want to work on X?" Additionally, a student asking another for help performing a task is excluded (such as asking another student how to sum a row in a spreadsheet, and the student telling them how).

In total, we found eight point events for inclusion from all ten videos. All such events were quick, directed comments related to a task the student was already engaging in. Therefore, as described in the main text, we conclude that no tasks were explicitly assigned by another student.

III. RESULTS

A. Identifying course-wide behavior patterns through cluster analysis

We analyzed the demographic composition of each behavior cluster by lab type (highly structured traditional or less-structured inquiry based), gender (students' self-reported gender identity of man or woman), and group composition (mixed-gender or single-gender groups). In all cases, when comparing the composition of behavior clusters, we used a chi-squared test of frequencies on the contingency tables of the raw counts.

When broken down by lab type [shown in Fig. 6(a)], 60% of the student profiles in the traditional labs were in the paper cluster, indicating that the majority of students in the traditional labs were high paper users. Students in the inquiry labs engaged in a more varied set of activities, demonstrated by the uniform distribution of student profiles across clusters. In the traditional labs, however, student profiles were predominantly found in the paper cluster, with few profiles in the remaining clusters.

Our data support the notion that labs with reduced structure provide a wider range of available roles. We tested this explanation by examining the range of roles within individual groups in each class type: Do members within a group predominantly fall into the same or different clusters? In the traditional labs, 43% of groups had all members in the same cluster (predominantly the paper cluster), whereas only 14% of groups in the inquiry labs had all members in the same cluster (Fig. 7).

We note that groups in the traditional and inquiry labs were of varying sizes. Groups in the traditional labs typically had three or four students, whereas groups in the inquiry labs typically had two or three members, with group sizes determined by logistical constraints of the lab

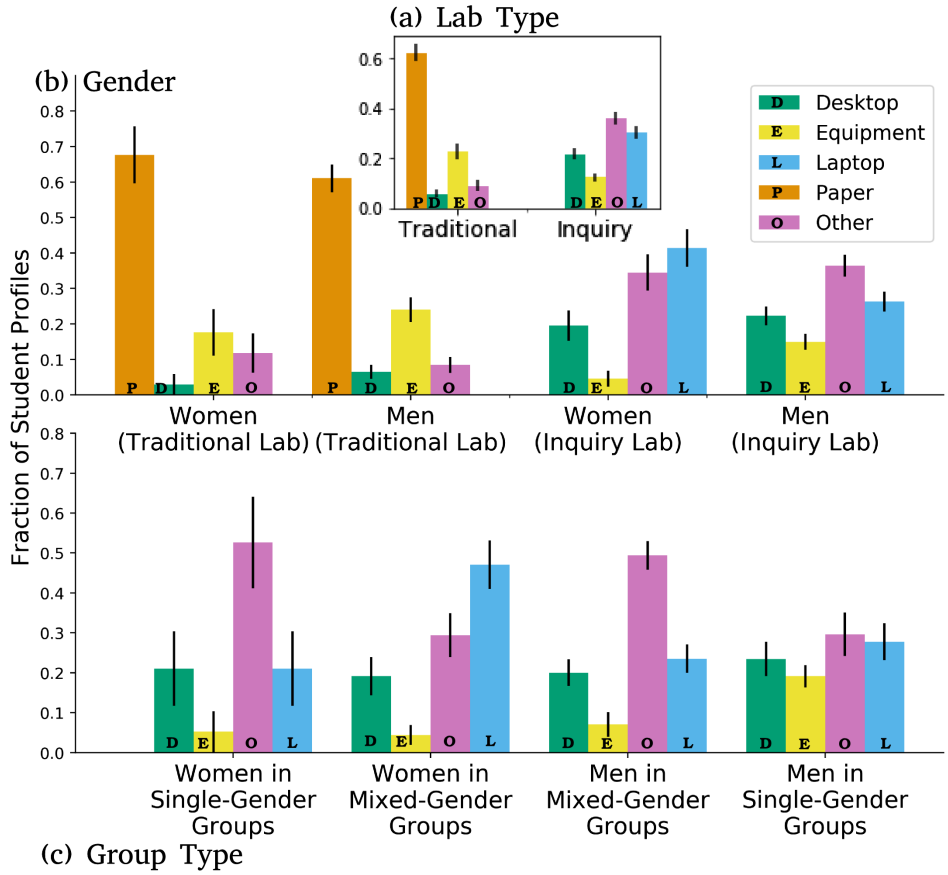

FIG. 6. Cluster compositions for each of the five clusters, broken down both by lab type, gender, and group composition. In all plots, the y axis represents a fraction of student profiles and errors are calculated using the standard error on the fraction of a population shown [see Eq. (1) for additional details]. (a) Cluster distributions broken down by lab type. (b) Clusters further broken down by gender. We see that there are disproportionately more women in the laptop cluster than men, and disproportionately more men than women in the equipment cluster. (c) Cluster distributions were further broken down in the inquiry lab by group type (men and women in mixed-gender groups and single-gender groups). Upon inspection, we see that the laptop difference remained, while a difference emerged in other. Furthermore, far more men are high-equipment users when in single-gender groups. Because of insufficient statistics, no comparison can be made with women in single-gender groups, and the data are presented for completeness.

spaces (such as the number of available lab benches given the size of each class) and mainly assigned randomly by the instructor. Moreover, mixed-gender groups also had between 1 to 3 women and 1 to 3 men. Observers documented the behavior of all students in every group, and kept track of which student was in which group. One could expect that, in groups with more members, there is an increased chance of task division occurring. While groups in the traditional labs typically had more members than those in the inquiry labs, Fig. 7 in fact shows proportionally fewer groups in the inquiry labs with members in identical clusters, supporting the conclusion that groups in the inquiry labs were more likely to divide tasks.

We infer that the set of available roles is much greater in the inquiry labs and that students assumed distinct roles from one another. The traditional labs were highly guided, leaving students little room for active decision-making about the experiment. While they worked in groups, each student was responsible for completing their own individual worksheet. As a result, the set of available roles was both confined and manifestly similar for all students. In contrast, the inquiry labs were designed to emphasize the process of experimentation and thus students supported in exercising agency for active decision making about the experiment. As a result, the set of available roles was larger and students could divide tasks in a variety of ways.

We next sought to evaluate whether men and women assume different roles. We decomposed the behavior clusters by gender and lab type, as shown in Fig. 6(b). Through a chi-squared test of frequencies, we found a statistically significant difference between men and women in the inquiry labs [$\chi^2(3) = 10.77$, $p = 0.01$, $V_{\text{Cramer}} = 0.15$], but none in the traditional labs [$\chi^2(3) = 3.27$, $p = 0.65$, $V_{\text{Cramer}} = 0.08$]. There were disproportionately more women in the laptop cluster than men and disproportionately more men in the equipment cluster than women.

Statistically significant differences also existed between men in mixed-gender versus single-gender groups, shown

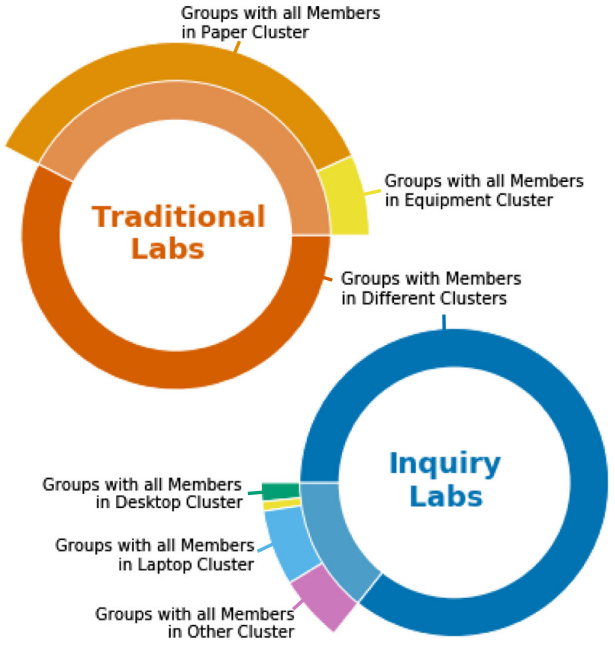

FIG. 7. Fraction of groups with members in identical clusters (light ring) and different clusters (dark ring) illustrating role division in the different labs. Almost half of groups in the traditional labs had all members in the same cluster (primarily paper cluster), whereas the majority of groups in the inquiry labs had members in multiple clusters indicating an increase in task division.

in Fig. 6(c) [$\chi^2(3) = 12.10$, $p = 0.007$, $V_{\text{Cramer}} = 0.15$]. When men were in single-gender groups, they were more likely to be in the equipment cluster and less likely to be in the other cluster than men in mixed-gender groups. Men in mixed-gender groups were more likely than their female group members to be in the other cluster, and women in mixed-gender groups were more likely than their male group members to be in the laptop cluster [$\chi^2(3) = 10.34$, $p = 0.02$, $V_{\text{Cramer}} = 0.15$]. Because of the small number of women in single-gender groups, we did not have statistical power to detect whether there are differences for women who were in mixed versus single-gender groups ($p > 0.17$ in all cases). Furthermore, due to insufficient statistics, we were unable to perform a similar analysis for groups of varying sizes.

The difference in men's behavior when in mixed-gender and single-gender groups may be indicative of the impact of social context on the roles students assume. In groups with only men, there may be different social dynamics compared to groups that include women, changing the set of available roles (and thus observed behaviors). For instance, the increased number of high-equipment users in men-only groups may be the result of "playfulness" [49] when women are not in the group, or that in mixed-gender groups members were more efficient with equipment use.

B. Quantifying the relative behaviors of students within groups

The cluster analysis in the previous section indicates that individual students took on different roles on a course-wide level but suggests that group composition may impact the group dynamics in a nontrivial way. To investigate roles within individual groups, and to ensure that different analysis methods obtain nonconflicting results, we compared each student's profile to those of their group members. We quantified the relative behaviors by constructing a deviating profile for each student to describe how they differed from their group's average profile (quantified as the numerical difference of the student profile from the group average, see Appendix for additional details). For example, if all students in a group behaved the same, the profiles of every student would match their group's average, and thus they would each have a deviation of zero for each code. The distribution of all students' deviations for each code has a mean of zero, as the deviations in every group must cancel each other out.

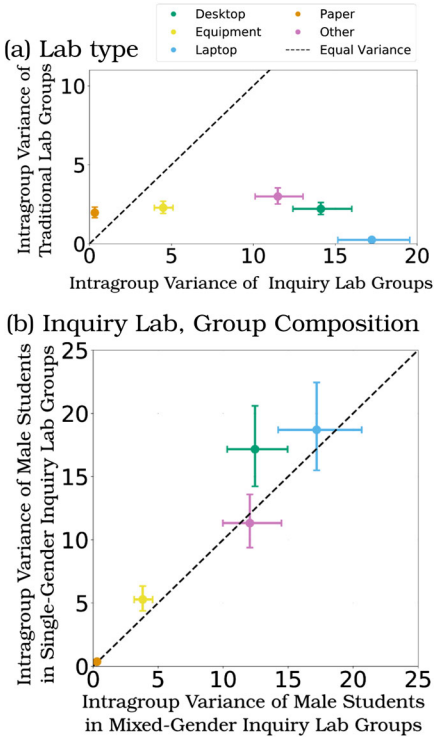

FIG. 8. Intragroup variances of the relative behaviors among students, signifying the amount of task division within groups. Each plot shows VAR(ΔN) for all student profiles contained within the labeled lab and group types along with their Bayesian confidence intervals. (a) Comparing across lab types, the intragroup variances are remarkably larger in the inquiry lab groups than in the traditional lab groups for all codes besides paper, indicating a greater range of behaviors and an increase in task division. (b) Within the inquiry labs, the intragroup variances are comparable among groups of differing composition suggesting that similar degrees of task division were taking place. (Female single-gender groups not included due to insufficient statistics).

However, the variances of these distributions (defined here as the intragroup variance) are not constrained and indicate the degree of task division. An intragroup variance of zero implies that any student's behaviors are completely indistinct from their group, while a large intragroup variance reveals a greater degree of divide and conquer.

In the traditional labs, the intragroup variance was very small for all coded behaviors other than paper [Fig. 8(a)]. This result supports the analysis and interpretation from the cluster analysis: groups in the traditional labs did not divide roles and each student behaved similarly to their group members. In the inquiry labs, intragroup variances were much larger for all codes apart from paper, which indicate a high degree of task division took place.

Within the inquiry labs, we found comparable intragroup variances among all coded behaviors regardless of the group's composition [that is, single-gender versus mixed-gender groups; Fig. 8(b)]. This result suggests that the group composition does not impact the group dynamics with respect to the amount of task division; that is, single- and mixed-gender groups divide roles to similar degrees.

However, within mixed-gender groups, these roles are divided along gender lines. The distributions of deviations for men and women in mixed-gender groups differed significantly for the laptop ($p = 0.001$) and other ($p = 0.009$) codes. Women handled a laptop or personal device more than their group members, and men participated in other activities more than their group members. Furthermore, men in mixed-gender groups appeared to handle equipment more often than the group average, however, this result was only marginally significant with the Bonferroni correction ($p = 0.012$). See Appendix and Fig. 9 for supporting data. To better understand these dynamics within mixed-gender groups, we sought a more fine-grained description of the roles students take on and how they are assigned.

C. Understanding specific student tasks and identifying role assignments

We captured video of a subset of individual groups for entire lab periods and identified the specific tasks associated with the coarse behaviors discussed in the previous sections. For example, when a student was handling the equipment, were they collecting data or setting up the apparatus? We identified the specific tasks through visual cues and students' speech. We then measured the total amount of time each student spent on each specific task.

The cluster and intragroup analyses found significant differences between men and women in mixed-gender groups with regards to laptop usage and other activities. The individual group video analysis found that women spent about twice as much time as men analyzing data on laptops ($14\% \pm 7\%$ of the lab period for women and $6\% \pm 3\%$ for men). However, we did not find a clear difference in the specific tasks associated with the other behavior between men and women in this subset of groups. The biggest difference among other tasks came from within-group behaviors such as talking, observing, or interacting with group members ($30\% \pm 4\%$ of the lab period for men and $26\% \pm 5\%$ for women).

We also used the single-group video analysis to identify that in almost all cases, students did not discuss the roles they would assume. Notably, there were no instances of explicit role allocation from peers in the group or from lab instructors. We conjecture students either self-assigned roles within groups, "fell into" roles, or directed each other through *positioning* (subtle verbal and nonverbal social cues [50,51]). Exploring mechanisms for role allocations is the focus of future study to better understand how roles become gendered. Tentatively, we conclude that the significant difference in roles is not the result of overt, explicit allocation. Rather, we infer that subtle interactions at the individual level accumulate to create class-level patterns.

IV. DISCUSSION AND CONCLUSIONS

In this study, we identified how student behaviors in a lab vary by lab type, gender, and group composition. From coarse-grained observations of what students were handling in the lab, we found that students in traditional labs generally behave similarly, spending most time writing on the lab worksheets. Behaviors in the inquiry labs were much more varied, with behaviors focused on using equipment and computers. Furthermore, women in the inquiry labs tended to be high laptop users (primarily analyzing data), while men were high equipment users (collecting data or manipulating the equipment). This pattern varied by group composition, however, where men in mixed-gender groups were much more often engaged in other behaviors (primarily talking to their peers), while men in single-gender groups were the high equipment users. Within-group analyses indicated that these differences were a result of group members taking on distinct roles, rather than whole groups tending towards similar behaviors. The role division was not a result of explicit allocation between group members.

Research indicates that providing students with more authentic lab experiences, often by removing structure to grant students more agency, improves student attitudes towards science and engagement in high-level scientific practices [39,52–55]. The results here suggest that by removing structure in labs, these curricula facilitate student-driven group work and open up a new set of group roles, but may unintentionally create inequitable learning environments or provide the opportunity for underlying inequities to manifest. Increased student agency, on its own, is insufficient for the creation of a supportive and equitable learning environment, where each student has the opportunity to freely pursue their own path in physics. Equitable participation must be actively built into curricula, to eliminate implicit inequities that can go on behind the scenes.

We have found that inquiry-based labs, designed to support student decision making, increased the variation in student behaviors when compared to the more traditional lab structure. Working collectively in groups, with a pedagogical structure that facilitated group work (such as having one electronic notebook per group as opposed to identical, individual worksheets) opened up new group roles and increased the range of behaviors students took on. Removing structure in lab activities so that students may take on a variety of roles supports a variety of students experiences during an activity. Through these experiences, we may communicate to students that there are multiple ways to contribute to science and to be a physicist.

However, the freedom for students to fall into roles without any guidance or pedagogical structure has the potential to introduce problematic inequities. While one could argue that allowing students to assume the roles they are more comfortable with may increase persistence in the course (regardless of whether or not they are gendered), we note that in the absence of structuring equitable participation and group work students may inadvertently fall back on cultural norms and expectations when taking on roles within their group and may rely on implicit biases when making these decisions. Each student's experience is unique in a classroom, but systematic differences in these experiences may have unintended, detrimental consequences. In this study, systematic gendered inequities (with men and women systematically taking on different group roles) and group behavior that depends on group composition (men behaving differently when in groups with other men versus when there is at least one woman) were statistically apparent only in a curriculum that provided ample agency. If such differences are supported in institutional settings, they can contribute to increased gender segregation through students' educational experience, and ultimately contribute to the large gender imbalance seen in the field as a whole.

The focus of this study was intentionally directed at students primarily intending to major in physics. While this narrow population limits generalization to nonphysics majors, it provides vital information with regards to group work, which has potential implications on students' identities as physicists and decisions to persist in physics [26]. This work also has implications for instruction. Our data, however, do not speak to the efficacy of different approaches at mitigating the issues observed. We can draw from previous literature to propose strategies that should be studied. For example, it has been shown that increased pedagogical structure combined with active learning can reduce the achievement gap in class work [56]. Therefore, actively building into lab curricula group roles (similar to those of cooperative grouping [57]) such as "group PI," "reviewer," or "science communicator" that have students actively think about how roles are assigned and make deliberate choices regarding role division could alleviate the unintended consequences of subconsciously acting on implicit biases, and is the focus of further research.

Previous work has identified many structural manipulations that support equitable participation in other learning environments [57,58]. Our results highlight that there may be unique challenges to equity in inquiry lab environments, where students divide roles associated with distinct experimentation tasks (such as analyzing data or handling equipment). The existence of role division is not inherently problematic. However, the different roles physics students take on can greatly influence their unique experience, identity formation, and sense of belonging, which, in turn, ultimately impact persistence and representation in the field [26,59,60]. With many calls to reform lab instruction to provide students with more authentic experiences and less structure, researchers have a responsibility to evaluate the potential side effects of such interventions. Given the many issues in representation and persistence in STEM, students' experiences should not be sacrificed for the increased learning benefits of these kinds of labs. Instructors have the responsibility of ensuring that the desired aspects of research and academia are being reinforced in these learning environments, and that we are not inadvertently reinforcing gendered roles by failing to actively intervene.

ACKNOWLEDGMENTS

We thank the teaching assistants and lab instructors for the course used in this study for their invaluable support and cooperation. We also thank Chris Gosling for valuable conversations and insight, and James Sethna for helpful feedback. This material is based upon work supported by the National Science Foundation under Grant No. 1836617, the President's Council for Cornell Women's Affinito-Stewart Grant, and the Cornell University College of Arts and Sciences Active Learning Initiative.

APPENDIX: STATISTICAL ANALYSIS OF INTRAGROUP VARIANCES

Here we present our intragroup analysis procedure to investigate whether roles emerged within individual groups. Each lab period involved groups of students working together as a team to progress through an experiment. We compared each student profile in a group to their group's average profile and quantified how a student deviated from their group's average for each code. Rescaling each profile in a group with respect to that group's average reveals the variations between the group-members' behaviors. We then compared whether there were any significant differences between the relative behaviors of men and women.

We quantified the relative behaviors of students by constructing each student's deviating profile. If the coded behaviors were distributed equally within a group, then the observations of each student would match the group's

average for each code. Denoting the observed count of a coarse behavior code for student S in group G with N^S_{code}, the expected count of that code for a student in that group is

$$\langle N_{\text{code}} \rangle_G = \frac{1}{M_G} \sum_{S \in G}^{M_G} N^S_{\text{code}}, \tag{A1}$$

where the sum runs over each student in one group with M_G total group members. From this expectation value, we calculate how student S deviates from their group's average

$$\Delta N^S_{\text{code}} = N^S_{\text{code}} - \langle N_{\text{code}} \rangle_G. \tag{A2}$$

These deviations reveal interesting behavior trends within groups. For instance, a student engaging in a particular task more than their group members would be revealed with a large and positive ΔN.

The distribution of deviations ΔN for each code provides information about task division within groups. When the distribution of deviations contains all group members, the mean is constrained to zero since each student's deviation cancels each other out by definition. However, the variance of these distributions for each code [the intragroup variance defined as $\text{VAR}(\Delta N)$] is not constrained and provides a measure of the amount of task division within a group. Zero variance among deviations would imply the students' behaviors are completely indistinct from another while a large variance would reveal a greater degree of divide and conquer.

In Fig. 8(a), we plot the intragroup variances for the two lab types. The relative behaviors within groups from the inquiry labs were highly varied when compared to the traditional labs, which exhibited remarkably less variance for all codes except paper. This result was confirmed with Levene's test to assess the equality of variances, where none of the p values from the test statistic for each code exceeded 10^{-5}. This disparity among intragroup variances was expected as the traditional labs were highly guided and students were required to fill out their own worksheet, while the inquiry labs were less guided and students were given more agency for active decision making about the experiment. The large intragroup variances in the inquiry lab groups signify a higher degree of task division taking place.

To investigate task division in the inquiry lab groups, we compared the intragroup variances for different group compositions [Fig. 8(b)]. We found comparable intragroup variances regardless of the group's composition for all behavior codes (every code's p value from Levene's test exceeding the $p = 0.01$ cutoff, with p values ranging from $p = 0.03$–0.9). The comparable intragroup variances signify that there was no significant difference in the degree of task division in the inquiry lab groups with respect to group composition.

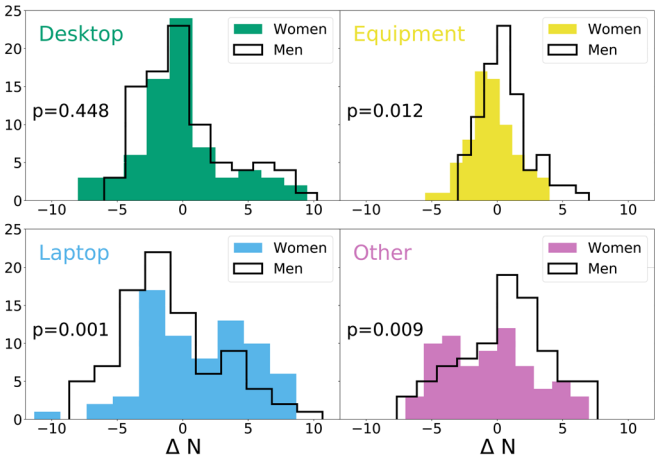

FIG. 9. Histograms of intragroup deviations for men and women within the inquiry lab's mixed-gender groups for the desktop, equipment, laptop, and other behavior codes, with the y axis representing the number of student profiles. Each student deviates from their group's average by ΔN [defined in Eq. (A2) in Materials and Methods A]. A positive ΔN denotes a student engaging in a behavior more often relative to their group members. We quote p values calculated from the Mann-Whitney U test statistic on all plots and find significant differences between men and women for the laptop and other codes. We also find a borderline result of men handling the equipment more than their group members.

To examine the relative behaviors of men and women, we shifted our focus to within mixed-gender groups. In Fig. 9, we plot the histograms of deviations for men and women in mixed-gender groups for the desktop, equipment, laptop, and other codes. We performed a Mann-Whitney U test as a nonparametric test to determine whether there were any significant differences among the distributions from men and women. We find significant differences in the deviations from men and women for the laptop ($p = 0.001$) and other ($p = 0.009$) codes. Women handled a laptop or personal device more than their group members, and men participated in other activities more than their group members. We also find that men in mixed-gender groups appear to handle equipment more often than the group average ($p = 0.012$), however, this result was only marginally significant with the Bonferroni correction.

[1] H. Pettersson, Making masculinity in plasma physics: Machines, labour and experiments, Science & Technology Studies **24**, 1 (2011).

[2] R. Scherr, Never mind the gap: Gender-related research in Physical Review Physics Education Research, 2005–2016, Phys. Rev. Phys. Educ. Res. **12**, 020003(E) (2016).

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

[4] S. Andersson and A. Johansson, Gender gap or program gap? Students' negotiations of study practice in a course in electromagnetism, Phys. Rev. Phys. Educ. Res. **12**, 020112 (2016).

[5] L. J. Sax, G. Holton, V. Van Horne, I. Nair, C. Davis, A. Ginorio, C. Hollenshead, B. Lazarus, and P. Rayman, Undergraduate science majors: Gender differences in who goes to graduate school, Rev. High. Educ. **24**, 153 (2001).

[6] S. L. Eddy and S. E. Brownell, Beneath the numbers: A review of gender disparities in undergraduate education across science, technology, engineering, and math disciplines, Phys. Rev. Phys. Educ. Res. **12**, 020106 (2016).

[7] K. Rosa and F. M. Mensah, Educational pathways of Black women physicists: Stories of experiencing and overcoming obstacles in life, Phys. Rev. Phys. Educ. Res. **12**, 020113 (2016).

[8] J. M. Nissen and J. T. Shemwell, Gender, experience, and self-efficacy in introductory physics, Phys. Rev. Phys. Educ. Res. **12**, 020105 (2016).

[9] 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).

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

[11] P. W. Irving and E. C. Sayre, Identity statuses in upper-division physics students, Cult. Stud. Sci. Educ. **11**, 1155 (2016).

[12] Z. Y. Kalender, E. Marshman, C. D. Schunn, T. J. Nokes-Malach, and C. Singh, Why female science, technology, engineering, and mathematics majors do not identify with physics: They do not think others see them that way, Phys. Rev. Phys. Educ. Res. **15**, 020148 (2019).

[13] P. W. Irving and E. C. Sayre, Becoming a physicist: The roles of research, mindsets, and milestones in upper-division student perceptions, Phys. Rev. ST Phys. Educ. Res. **11**, 020120 (2015).

[14] E. Wenger, R. A. McDermott, and W. Snyder, *Cultivating Communities of Practice: A Guide to Managing Knowledge* (Harvard Business Press, 2002) p. 284.

[15] A. J. Gonsalves, A. Danielsson, and H. Pettersson, Masculinities and experimental practices in physics: The view from three case studies, Phys. Rev. Phys. Educ. Res. **12**, 020120 (2016).

[16] B. Francis, L. Archer, J. Moote, J. DeWitt, E. MacLeod, and L. Yeomans, The construction of physics as a quintessentially masculine subject: Young people's perceptions of gender issues in access to physics, Sex Roles **76**, 156 (2017).

[17] S. L. Li and D. Demaree, Assessing physics learning identity: Survey development and validation, AIP Conf. Proc. **1413**, 247 (2012).

[18] J. Stake and S. Nickens, Adolescent girls' and boys' science peer relationships and perceptions of the possible self as scientist, Sex Roles **52**, 1 (2005).

[19] E. W. Close, J. Conn, and H. G. Close, Becoming physics people: Development of integrated physics identity through the Learning Assistant experience, Phys. Rev. Phys. Educ. Res. **12**, 010109 (2016).

[20] R. Lock, J. Castillo, Z. Hazari, and G. Potvin, *Proceedings of the 2015 Physics Education Research Conference, College Park, MD* (AIP, New York, 2015), pp. 199–202.

[21] Z. Hazari, E. Brewe, R. M. Goertzen, and T. Hodapp, The importance of high school physics teachers for female students' physics identity and persistence, Phys. Teach. **55**, 96 (2017).

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

[23] American Association of Physics Teachers, AAPT Recommendations for the Undergraduate Physics Laboratory Curriculum, Tech. Rep. (American Association of Physics Teachers, College Park, PA, 2014), https://www.aapt.org/Resources/upload/LabGuidlinesDocument_EBendorsed_nov10.pdf.

[24] N. W. Brickhouse, Embodying science: A feminist perspective on learning, J. Res. Sci. Teach. **38**, 282 (2001).

[25] L. Archer, J. Moote, B. Francis, J. DeWitt, and L. Yeomans, The "Exceptional" physics girl: A sociological analysis of multimethod data from young women aged 10–16 to explore gendered patterns of post-16 participation, Am. Educ. Res. J. **54**, 88 (2017).

[26] H. B. Carlone and A. Johnson, Understanding the science experiences of successful women of color: Science identity as an analytic lens, J. Res. Sci. Teach. **44**, 1187 (2007).

[27] J. Butler, *Gender Trouble: Feminism and the Subversion of Identity* (Routledge, New York, 1999).

[28] 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).

[29] 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).

[30] 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).

[31] 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).

[32] 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).

[33] N. G. Holmes, I. Roll, and D. A. Bonn, Participating in the physics lab: Does gender matter?, Phys. Canada **70**, 84 (2014).

[34] 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).
[35] M. Laeser, B. M. Moskal, R. Knecht, and D. Lasich, Engineering design: Examining the impact of gender and the team's gender composition, J. Engin. Educ. **92**, 49 (2003).
[36] 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).
[37] C. Gosling, Identity as a research lens in science and physics education, J. Belonging Identity Language Diversity **1**, 62 (2017).
[38] 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).
[39] N. G. Holmes, C. E. Wieman, and D. A. Bonn, Teaching critical thinking, Proc. Natl. Acad. Sci. U.S.A. **112**, 11199 (2015).
[40] 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).
[41] N. G. Holmes and E. M. Smith, Operationalizing the AAPT learning goals for the lab, Phys. Teach. **57**, 296 (2019).
[42] N. G. Holmes, B. Keep, and C. E. Wieman, Developing scientific decision making by structuring and supporting student agency, Phys. Rev. Phys. Educ. Res. **16**, 010109 (2020).
[43] J. H. Corpus and S. V. Wormington, Profiles of intrinsic and extrinsic motivations in elementary school: A longitudinal analysis, J. Exp. Educ. **82**, 480 (2014).
[44] J. A. Schmidt, J. M. Rosenberg, and P. N. Beymer, A person-in-context approach to student engagement in science: Examining learning activities and choice, J. Res. Sci. Teach. **55**, 19 (2017).
[45] J. A. Hartigan and M. A. Wong, Algorithm AS 136: A K-means clustering algorithm, J. R. Stat. Soc. Series C (Applied Statistics) **28**, 100 (1979).
[46] R. L. Thorndike, Who belongs in the family?, Psychometrika **18**, 267 (1953).
[47] L. v. d. Maaten and G. Hinton, Visualizing data using t-SNE, J. Machine Learn. Res. **9**, 2579 (2008).
[48] O. Friard and M. Gamba, BORIS: a free, versatile open-source event-logging software for video/audio coding and live observations, Methods Ecol. Evol. **7**, 1325 (2016).
[49] C. Hasse, Learning and transition in a culture of playful physicists, Eur. J. Psychol. Educ. **23**, 149 (2008).
[50] B. Davies and R. Harré, Positioning: The discursive production of selves, J. Theory Soc. Behav. **20**, 43 (1990).
[51] M. Berge and A. T. Danielsson, Characterising learning interactions: A study of university students solving physics problems in groups, Res. Sci. Educ. **43**, 1177 (2013).
[52] B. R. Wilcox and H. J. Lewandowski, Open-ended versus guided laboratory activities:Impact on students' beliefs about experimental physics, Phys. Rev. Phys. Educ. Res. **12**, 020132 (2016).
[53] E. Etkina, A. Karelina, M. Ruibal-Villasenor, D. Rosengrant, R. Jordan, and C. E. Hmelo-Silver, Design and reflection help students develop scientific abilities: Learning in introductory physics laboratories, J. Learn. Sci. **19**, 54 (2010).
[54] S. E. Brownell, M. J. Kloser, T. Fukami, and R. Shavelson, Undergraduate biology lab courses: Comparing the impact of traditionally based "Cookbook" and authentic research-based courses on student lab experiences, J. Coll. Sci. Teach. **41**, 36 (2012).
[55] D. J. Adams, Current trends in laboratory class teaching in university bioscience programmes, Biosci. Educ. **13**, 1 (2009).
[56] D. C. Haak, J. HilleRisLambers, E. Pitre, and S. Freeman, Increased structure and active learning reduce the achievement gap in introductory biology, Science **332**, 1213 (2011).
[57] P. Heller and M. Hollabaugh, Teaching problem solving through cooperative grouping. Part 2: Designing problems and structuring groups, Am. J. Phys. **60**, 637 (1992).
[58] K. D. Tanner, Structure matters: Twenty-one teaching strategies to promote student engagement and cultivate classroom equity, Cell Biol. Educ. **12**, 322 (2013).
[59] K. Rainey, M. Dancy, R. Mickelson, E. Stearns, and S. Moller, Race and gender differences in how sense of belonging influences decisions to major in STEM, Int. J. STEM Educ. **5**, 10 (2018).
[60] A. J. Fisher, R. Mendoza-Denton, C. Patt, I. Young, A. Eppig, R. L. Garrell, D. C. Rees, T. W. Nelson, and M. A. Richards, Structure and belonging: Pathways to success for underrepresented minority and women PhD students in STEM fields, PLoS One **14**, e0209279 (2019).