Gender inequity in individual participation within physics and science, technology, engineering, and math courses

Alessandra M. York,[1,*] Angela Fink,[2,†] Siera M. Stoen,[3,‡] Elise M. Walck-Shannon,[4,§] Christopher M. Wally,[5,‖] Jia Luo,[6,¶] Jessica D. Young,[7,**] and Regina F. Frey[8,††]

[1]*Department of Biology (CIRCLE), Washington University in St. Louis, St. Louis, Missouri 63130, USA*
[2]*(CIRCLE), Washington University in St. Louis, St. Louis, Missouri 63130, USA*
[3]*Department of Physics (CIRCLE), Washington University in St. Louis, St. Louis, Missouri 63130, USA*
[4]*Department of Biology, Washington University in St. Louis, St. Louis, Missouri 63130, USA*
[5]*Department of Biology (CIRCLE), Washington University in St. Louis, St. Louis, Missouri 63130, USA*
[6]*Department of Chemistry, Washington University in St. Louis, St. Louis, Missouri 63130, USA*
[7]*Department of Chemistry, University of South Florida, Tampa, Florida 33620, USA*
[8]*Department of Chemistry, University of Utah, Salt Lake City, Utah 84112, USA*

(Received 11 June 2021; accepted 28 October 2021; published 2 December 2021)

Gender inequities continue to persist within science, technology, engineering, and mathematics (STEM) disciplines, even at the undergraduate level. This has led researchers to further examine potential factors that contribute to retention and persistence of undergraduates in STEM fields. In this study using classroom observations, we examined gender equity in individual verbal participation in large introductory physics courses, and compared our results to observations in introductory courses in other STEM disciplines. We found that in introductory physics courses, men had disproportionately higher rates of individual verbal participation than women. Observation-level analysis confirmed that three-quarters (76.2%) of the physics observations had descriptively higher than expected participation by men and almost a quarter of observations (23.8%) were statistically significant for a gender imbalance in individual verbal participation. We then sought to determine if any pedagogical strategies or student behaviors correlated with a more equitable classroom to better understand what drives gender inequity in participation, and found three classroom behaviors—an increasing amount of instructor questions, group responses from students, and student questions—correspond with a more gender equitable classroom. Student-level survey data, which mirrors the observation data, also show that self-reported levels of individual participation have small, significant correlations with both course-level belonging and inclusivity. The introductory physics results were mostly replicated in the other STEM disciplines, despite their differences in course structure. The patterns of individual participation were still disproportionately higher for men, with two-thirds of observations displaying a bias towards more men participating. Student-level survey data continued to mirror the observation data, and small, significant correlations between student self-reported participation and course-level belonging and inclusion were found. However, only the number of student questions correlated with a more equitable classroom in other STEM courses. This study extends the conversation on the relationship between active learning and equity in the classroom, demonstrating a need to move beyond mere inclusion of active pedagogies towards proactive facilitation of equitable and comfortable verbal participation by all students. Practical strategies for encouraging inclusive classroom dialogue, such as transparency, growth-mindset messaging, and multiple modes of engagement, are discussed.

DOI: 10.1103/PhysRevPhysEducRes.17.020140

*amyork@wustl.edu
†amfink@wustl.edu
‡sstoen@wustl.edu
§ewalck-shannon@wustl.edu
‖cwally@wustl.edu
¶jluoa@wustl.edu
**jdyoung2497@gmail.edu
††Corresponding author.
gina.frey@utah.edu

I. INTRODUCTION

While many strides have been made within the fields of science, technology, engineering, and mathematics (STEM), we still do not observe gender equity. Many STEM disciplines exhibit an underrepresentation of women at the undergraduate level [1] that becomes further exacerbated at higher levels within academia, creating what has been called the "leaky pipeline" [2]. In the field of physics, those who advance to higher academic levels see even greater gender disparities. For example, while men and women tend to have similar numbers of promotions and publications 10–15 years after graduation, data show that the time for

promotion to full professor is one year longer for women than men [3]. Men are also 45% more likely to present at a conference as an invited speaker, 27% more likely to act as a journal editor, and 53% more likely to have enough funding for their research. Porter and Ivie's study showed that, in addition, women report their careers are more impacted by family demands than men, contributing to discrepant professional opportunities between men and women in the field [3].

To understand the persistence of these gender inequities, the first step many educators and administrators pursue involves investigating how to increase retention of women undergraduates in STEM. The number of women who intend to pursue STEM majors when entering higher education is nearly as high as the number of men [4], yet they are disproportionately more likely to transfer to a non-STEM major than men [5]. A similar pattern is seen in the transfer rate of women versus men out of physical sciences or quantitative fields. For example, Turnbull *et al.* [6] found that women physics undergraduates are more likely than men physics undergraduates to transfer into life science fields, which now award more bachelor's degrees to women than men [1]. We hypothesize that inequitable participation may contribute to undergraduate women leaving STEM fields at a higher rate than undergraduate men, as participation has been shown to be a predictor of performance [7] and a major factor in retention [5]. As a first step towards testing that hypothesis, the current study aims to assess the prevalence of gender inequities in the classroom participation in physics and other STEM disciplines. Specifically, this study explores, via classroom observations, individual verbal participation patterns in large introductory STEM courses (introductory physics, introductory biology, general chemistry, and calculus).

A. Gender disparities in physics courses

The literature has explored if gender disparities are present in a range of outcomes in undergraduate physics courses. For instance, social psychological measures like self-efficacy, social belonging, and STEM identity frequently reveal gender differences in students' subjective experiences of physics classrooms. Men tend to have higher levels of self-efficacy in physics [8–10], while differences in self-efficacy are absent from other introductory STEM courses, non-STEM courses, or outside activities [11]. Shaw [12] observed a significant difference in men having higher self-efficacy than women in a physics course for non-majors, with a trend in the same direction emerging in an algebra-based major-level course. Following the same pattern, men also report a higher sense of belonging in physics [13,14] and have stronger STEM identities; i.e., they are more likely to perceive themselves as a "physics person" than are women [14,15].

In contrast to the clear evidence that men and women experience physics courses differently, the literature exploring performance differences between men and women in STEM lacks consistency [16–18]. For example, both Tai and Sadler [18] and Miyake *et al.* [17] document achievement gaps between genders in calculus-based introductory physics courses, while Lauer *et al.* [16] find no achievement gap in their calculus-based introductory physics course. This equivocal evidence could be due to various factors, such as how students self-select the courses they take or the use of different outcome variables to index performance (i.e., exam performance versus overall course grade) [19].

In addition, studies focusing on conceptual learning instead of course performance suggest both gender differences and similarities. When examining physics concept inventories, men have been reported to score higher on the Force Concept Inventory (FCI) than women [14,20–23]. However, Seyranian *et al.* [14] found that although men entered the course with higher overall physics knowledge, there was no statistical difference between genders for gains in FCI score over the course of the semester. Given that men and women acquire physics knowledge at similar rates across the semester, Seyranian *et al.* [14] concluded that the social psychological experiences of women in the classroom may be the key difference between genders in physics. The current study investigates the role classroom participation plays in these disparate experiences.

B. Participation and gender

The literature that examines verbal participation and gender in STEM classrooms is relatively new and mostly concentrated in biology courses. Eddy *et al.* [24] examined gender gaps in achievement and participation in introductory biology courses at a large state university. They recorded the perceived gender of individual participants in the 23 classes they observed, and found that, while women made up approximately 60% of students in these courses, they answered less than 40% of instructor-posed questions to the class. The study examined different types of individual verbal participation and consistently found patterns of underrepresentation of individual women participating across the board. The only participation methods where gender differences were not observed were random call and spontaneous student questions. Similar to Eddy *et al.* [24], Aguillon *et al.* [25] examined large introductory biology courses for gender-specific differences in participation. They observed the perceived gender of participants in an introductory biology course over multiple semesters and found men participated more than expected based on class composition in most of the seven participation categories they examined, especially in voluntary responses. They also found women in the class identified more strongly with their gender than men, and were more likely to report that other people would judge them based on their gender, perhaps due to stereotype threat [26]. Bailey *et al.* [27] also conducted observations recording the perceived gender of individual

participants in 34 life science courses at a large university. Their findings echoed the studies above of men participating more than expected, but found these gaps could vary based on the course observed. They found when using mixed linear models to predict the gender participation gaps, the two characteristics that predicted higher rates of participation among females was in female-majority classrooms or classes where the instructor called on the majority of raised hands.

Other work has focused on the student perception of individually participating in class and why they may not participate. In a follow-up study using self-reported surveys, Eddy *et al.* [28] found a gender difference in anxiety towards some types of participation, but not towards other types of participation. For example, they determined that women disproportionately report having more anxiety than men report having in whole-class discussions in their large course, but an equally low level of anxiety as men in peer discussions. Nadile *et al.* [29] examined student perceptions of asking and answering questions in large-enrollment science classes, and found women were 2.4 times more likely than men to report never asking questions, and were 2.8 times more likely to report feeling uncomfortable asking questions themselves. They also were 3.9 times more likely to describe feelings of anxiety when asking questions, and were nearly 2 times more likely to report that they did not feel they knew the material well enough to ask a question.

Although not yet studied in connection with participation in class, sense of belonging is determined by a person's self-reported answers to questions such as "Do I feel comfortable with my classmates?" and "Do I feel comfortable with my instructor"; therefore, course-level sense of belonging describes a student's sense of their interpersonal relationships in the course and their level of comfort and confidence in the course. Studies have shown that affective variables such as belonging (as defined by Walton and Cohen [30,31]) and inclusion in a course affect student success in their STEM courses (i.e., course performance and persistence in the series) [32,33]. Both Edwards *et al.* [32] and Fink *et al.* [33] found both lower levels of course-level social belonging and higher levels of course-level belonging uncertainty for women than men in general chemistry courses. Rainey *et al.* [34] showed that sense of belonging affected whether students remained in STEM majors, and that women and persons of color had lower sense of belonging than white men. As seen in these studies, there is increasing evidence that belonging affects student performance and retention in STEM, however, much less is known about the mechanisms(s) of belonging on STEM performance. The model of Zumbrunn *et al.* [35] found that a supportive classroom environment can improve student belonging, affecting a student's level of engagement, motivation, and finally achievement in the course. So far in the literature, course-level inclusion has been typically looked at through the lens of reducing stereotype threat [13,36], rather than by seeking the answers to what makes an inclusive classroom environment. Participants with a higher sense of belonging in math tend to feel less anxious about math and more confident in their abilities [36]. Mentioned in the prior section, Stout *et al.* had similar findings to Good *et al.*, where stereotype threat negatively predicted women's sense of belonging in physics courses [13]. These studies show the importance of belonging to performance and retention, and that classroom environment and course inclusivity affect student belonging. We, therefore, hypothesize and test in this study that course-level belonging and inclusivity are potential reasons why women may be less likely to verbally participate than men in STEM courses, specifically in physics courses.

Few studies examine verbal participation and gender in STEM disciplines outside of biology, or look across STEM disciplines. Ballen *et al.* [37] conducted classroom observations using predicted gender as a proxy to examine gender gaps in participation across six different universities of various sizes and four STEM disciplines. They examined six different hypotheses to potentially explain gender gaps in individual verbal participation: abundance of student-instructor interactions per class period, diversity of teaching strategies, instructor gender, proportion of women in the class, class size, and whether the course was lower or upper division. When examining whether the gender equity of participation (indexed by likelihood ratios) was predicted by using those six variables, they found smaller class sizes and instructors using a diverse set of teaching strategies created more equitable participation across genders. They did not disclose whether participation differed across discipline.

In the broader literature examining gender and participation across STEM and non-STEM disciplines, the results appear mixed: some studies found women have higher levels of participation than men [38–40], some found that men have higher levels of participation than women [24,41,42], and others found no statistical difference between genders [43–45]. Given these inconsistent results, more research needs to be conducted to understand how different classroom-level and student-level variables influence participation dynamics in the classroom. For instance, the subject being taught can have an effect on participation levels, with science courses showing less participation than other subjects like arts and humanities [41,46]. Classroom size [46], course level [39], course structure [47], student age [39], student fears of criticism and peer disapproval [48], and student confidence levels [48] all contribute to varying participation levels. Results also may depend on the types of evidence used to gauge participation levels. Karp and Yoels [49] found that over 90% of students perceived no difference in men or women's participation levels, but when observed, men were more likely to participate, especially in classes taught by men. Therefore, students may not be conscious of the participation disparities in the classroom, but nevertheless

these participation disparities may contribute to undergraduate women's experience and decisions to leave STEM fields at a higher rate than undergraduate men.

Participation is important to examine because research has illustrated many benefits for students who participate in undergraduate classrooms. Students who verbally participate more have higher levels of motivation [50], learn the material better [48,51,52], are stronger critical thinkers [53,54], engage in higher levels of learning because they memorize less [55], and earn higher grades [7]. Students who participate also tend to improve in skills valuable outside of the classroom such as communication skills [56,57], group interactions [58], and functioning in a democratic society [59]. The fact that participation is valuable to students would not surprise them; research shows students are aware that participation is valuable for their learning [39]. However, students also appear to have different definitions of "participation" than their instructors [39], which may contribute to the finding that students report higher levels of participation than instructors [60]. Some students in Fritschner's study defined participation as attendance, active listening, doing assignments or being prepared for class, whereas faculty were more likely to describe participation as asking or answering questions, or other verbal interactions [39].

Despite the clear benefit of in-class participation, research consistently shows that only a handful of students in any given classroom actually participate on a regular basis [38,39,41,49,61,62]. Karp and Yoels [49] described this phenomenon as the "consolidation of responsibility," where the majority of students feel they can be passive observers (or participate only on occasion). In the current study, we explore the possibility that this small group of active participants is disproportionately men in our large introductory STEM courses and explore what factors may influence gender equity in the classroom.

C. THE CURRENT STUDY

Our study begins by exploring individual verbal participation patterns in calculus-based introductory physics classrooms to establish whether gender equitable participation is observed, and then compare those results to other introductory STEM classes. Given the value of classroom participation, we examine whether there are gender gaps in individual verbal participation, and if so, what instructor or student behaviors correlate with those patterns. We achieve these goals using a classroom observation protocol called observation protocol for active learning, or OPAL [63], and a new, accompanying tool called the gender participation map. This map allows us to document the location and, following previous research [24,27,37], perceived gender of students who verbally participate during the classroom observation. Although student participation can take many different forms in the classroom [57], this study focuses on students who individually and voluntarily contribute to the discussion in class. In addition, these observation data are compared with student self-reported survey data regarding their own participation, and feelings of social belonging and inclusion in the course.

Study 1 was guided by the following research questions about individual student verbal participation in introductory physics I and II:

- Is individual verbal participation in the introductory physics courses gender balanced?
- Are more interactive classes correlated with more equitable individual verbal participation? Are there any instructor or student behaviors that are correlated with equal participation?
- Is individual participation via clickers equitable?
- Do student perceptions of their individual verbal participation in the introductory physics class mirror the observational data?
- Does self-reported verbal participation correlate with course-level sense of belonging and inclusivity at the end of the course?

In study 2, we examined whether the patterns observed in introductory physics are replicated in other introductory STEM classes (introductory biology I and II, general chemistry I and II, and calculus I and II), which had different amounts and patterns of active learning, different class sizes, and different proportions of men and women.

II. METHODS

A. Study setting

Data were collected at a midsized, private, selective research institution in the Midwestern United States during the academic year 2018–2019 and fall semester of 2019. Following receipt of a grant (given in 2013) from Association of American Universities' (AAU) undergraduate STEM Education Initiative, which aimed to increase the use of active learning and other evidence-based instructional practices in the large lower-level STEM courses, all students were provided access to use clickers, a type of student response system in their classes, free of charge. This led to increased usage of active-learning strategies in all introductory science and mathematics courses. In 2017, the institution was awarded an HHMI Inclusive Excellence grant, which focused on increasing the inclusivity in these large introductory STEM courses and led us to examine the equity of participation in these courses.

The majority of observations were conducted in an introductory calculus-based physics course series consisting of multiple sections taught by different instructors; however, all sections followed the same content, policies, and assessments. The introductory physics course was designed differently than the other introductory math and science courses at this institution, because twice a week students were asked to complete videos through FlipItPhysics to gain some of the content knowledge before attending class. Actual lecture

time was used to go over more challenging material and examples for students to problem solve together. Individual PowerPoint slides varied among the instructor team, but the same topics were covered at the same time and every instructor presented a core set of the same example problems (around one-quarter of all lecture examples) to enhance the similarity across sections. Demonstrations were also frequently used, where students would predict and then discuss their observations. Starting in Fall 2019, the lectures from every instructor were recorded and available for students to watch, of which many students took advantage.

For academic year 2018–2019, assessments included participation in class via clicker activities (3%), FlipItPhysics lecture videos (5%), weekly homework assignments through Mastering Physics (15%), exams (57%), and lab (20%). In fall 2019, the course structure was adjusted because lab became a separate course. FlipItPhysics lecture videos (5%), clicker activities (5%), homework through Mastering Physics (10%), and exams (60%) were still being used in the course, but the points originally going to lab were now assigned for instructor-written challenging problems via Gradescope (20%). However, the balance between high and low-stakes assignments stayed the same across the three semesters observed. Information about the other STEM courses is provided in the Supplemental Material [64].

B. Collection of quantitative observation data

1. Sample

Participating instructors (physics = 7, other STEM = 15) were full-time faculty, consisting of eight tenured, four tenure track, and nine non-tenure track faculty. These instructors were drawn from four (of six total) STEM departments in the College of Arts and Sciences on campus (i.e., Biology, Chemistry, Mathematics, and Physics). These introductory courses were targeted for the study because of their participation on the AAU and HHMI grants, and they engage the majority of undergraduate STEM or pre-health students. All instructors of the observed courses participated. In introductory physics, there were 7 instructors (6 men and 1 woman), comprising 3 non-tenure track, 3 tenure track, and 1 tenured faculty.

All the observed science and math courses can be categorized as lecture based for two reasons. First, the university describes these courses as lectures in official documents. Second, our observational data indicate that the dominant teaching strategy used in every class session was lecture. It should be noted that many of the courses included additional components such as laboratories or recitations, but only the lecture portion of each course was observed. The individual lecture sections had enrollments ranging from 70 to 338 students (average section enrollment: Physics, $N = 96$; other STEM courses, $N = 194$).

All faculty members were observed approximately 3 times during a given semester. Thirty-two percent ($n = 7$) of the 22 faculty participants were observed during multiple semesters. If an instructor taught multiple sections of the same course in a semester, we tried to observe the section with the larger enrollment. When we observed the same course over multiple semesters, we observed the same topics each semester to control for variation in content that may be more or less amenable to active learning [65,66]. Data were collected for 1.5 academic years, from fall 2018 through fall 2019.

2. Tools

We used the observation protocol for active learning (OPAL) [63] to document our observations in each course. Similar to COPUS [67] and TDOP [68], OPAL captures both instructor and student behaviors in 2-min intervals. To study individual verbal participation, we created a gender participation map to supplement the OPAL observation tool and documented individual verbal participation of students (Supplemental Material [64]). For each 2-min time interval, the gender participation map divides the classroom into six quadrants (front right, front center, front left, back right, back center, and back left), and the observer records the perceived gender of any individual student participant within that time interval. We should note that following previous research [24,27,37], the perceived gender of an individual participant is decided by the observer, and may not be the gender that the student identifies with; therefore, it is an imperfect variable to study gender disparities in the classroom. If the observer was unsure of the participant's gender, they used "U" for unknown. Less than 1% (7/1144) of individual participants were coded "unknown," and these participants were removed from the study. At the end of each observation, we summed the total number of individual participants and calculated the percentages of perceived men and women who individually participated verbally. An approximate number of students in each of the six quadrants was also recorded, but the location data are not reported in this study. Importantly, the map does not keep track of individuals who participate more than once, due to observers' inability to identify individual students in large enrollment classes where seats are not assigned. Instead, each instance of an individual participating is coded as a unique participant, and repeat participants may therefore influence the gender equity of observations. Implications of this data collection method are discussed in Sec. V.

3. Timelines

The quantitative OPAL data from each observation (physics, $N = 42$; other STEM courses, $N = 54$) were transformed into OPAL timelines (see the Supplemental Material in Frey *et al.* [63]). Seven observations were eliminated because they had no individual participants; one was eliminated due to the lack of a gender participation

TABLE I. Example OPAL codes in each segment type*.

Segment label	OPAL codes for instructor	Brief description for instructor	OPAL codes for students	Brief description for students
Lecture	LHV	Lecture with handwritten visuals	L_I	Listening to instructor
Clicker activity	PSb	Posing problem-solving activity on the board	QG	Discussing question in groups
	Sfu	Summary follow-up	V_T	Voting with technology
Demonstration or video	PDV	Passive demonstration or video	L_I	Listening to instructor
	Dfu	Discussion follow-up		
Other active learning	PSv	Posing problem-solving activity verbally	Ind	Thinking or working individually
	Dfu	Discussion follow-up		

*This is not the exhaustive list of all the codes that can fall into a segment, only a subset of the most common codes used. For a full code list with detailed descriptions, please see the Supplemental Material of Frey *et al.* [63].

map. Following the procedure used in Solomon *et al.* [69], we separated the quantitative timeline data into segments corresponding to the teaching strategies used: lecture, clicker activities, demonstrations or videos, and other active-learning activities [(OAL); in this study, primarily small-group or individual problem solving without the use of clickers]. One additional observation was excluded from analysis because it consisted of lecture only (i.e., no active learning was observed) and therefore would not fit into our active-learning framework. The observations that were eliminated from the study were all from the other STEM courses observed; thus, the final sample included 87 observations across four STEM disciplines (physics, $N = 42$; other STEM courses, $N = 45$).

While almost all the quantitative timeline data could be categorized in one of these four teaching strategies, there were a limited number of intervals we labeled "not coded." These intervals were often course announcements at the beginning or end of the class session, and they amounted to only 3.9% of the class session on average. Occasionally, different segments (e.g., lecture and clicker activities) would overlap within one two-minute interval due to the ability for multiple classroom activities to occur during the same time interval. In those cases, we divided the interval in half, attributing one minute to each of the teaching strategy segments, following the procedure in Solomon *et al.* [69]. When applying the segmenting technique, the codes in each two-minute interval were our guide in determining where one segment began and another ended (see the Supplemental Material in Frey *et al.* [63] for both the codes and guidance on how to segment). Example OPAL codes corresponding to each of the four segment types are listed in Table I and samples of segmenting for each subcategory are provided in Fig. 1. The categories are discussed below in Sec. III A and are found in Table II (detailed descriptions in Supplemental Material [64]). Additional information, such as complete sample timelines for each subcategory and how we applied the segmenting, can be found in the Supplemental Material [64].

C. Visual analysis of observation data

Following the procedure in Solomon *et al.* [69], once all the timelines were generated and segmented, we conducted a visual analysis by comparing observation timelines within and between faculty. Qualitative-analysis methods, including visual analysis, are typically iterative in nature, meaning that the analyses are conducted in multiple steps before arriving at the final results, and multiple research members are involved in this visual analysis [70]. Qualitative results are typically in the form of categories, sometimes subdivided into narrower subcategories, that are then described in detail with samples of data provided. Such results can then be used to calculate frequencies for quantitative analysis, which we do in this study.

The visual analysis in this study began with one research-team member (A. Y.) visually inspecting every timeline and looking for similarities and differences in the implementation (e.g., timing, duration, and frequency) of the four teaching strategies described above, similar to Solomon *et al.* [69]. A. Y. determined that some timelines did not align well with the original framework; the implementation of active learning had evolved or become more complex since the previous study. In such cases, A. Y. took notes about the additional pedagogical strategies and parameters used to differentiate the timelines; for example, noting a contrast between instructor versus student-prompted interaction.

D. Categorization of class sessions based on observation data

To determine the final categorization of each observation, two additional research-team members (A. F. and R. F.) independently visually inspected all timelines and categorized them into the original categories, separating out those that did not fit well. Any discrepancies among the three categorization decisions by A. Y., A. F., and R. F. were resolved through detailed and critical discussion. Interrater reliability was not calculated due to the modest

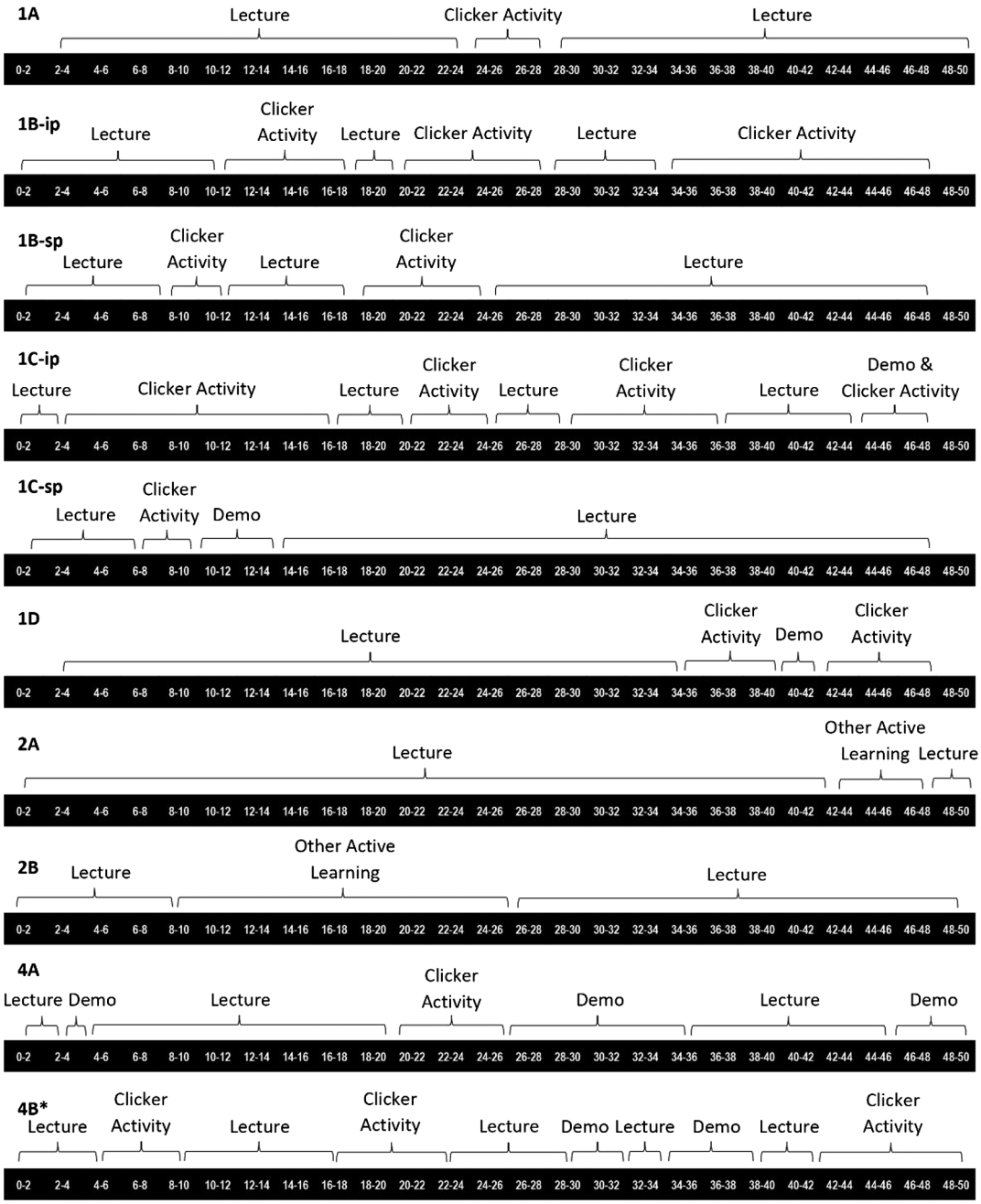

FIG. 1. Sample segmenting for each subcategory. In the top left corner, the subcategory the timeline represents is noted (see list of categories and subcategories in Table II). The center black bar shows the 2-min time intervals during an observation. The brackets indicate the teaching strategy used in that time interval. The combined "Demo & Clicker Activity" brackets are used to visualize clicker activities based on a demonstration, but the time spent on each individual teaching strategy was noted during the observation and used in our analyses. The asterisk on the last timeline denotes that this timeline shows the first 50 min of an 80-min observation. The corresponding full timelines for each segmented sample in this figure are in the Supplemental Material [64].

sample size, and hence analysis triangulation (between the three researchers) was used and we reached consensus on all observations [70,71]. Each team member provides a unique perspective on the data: (1) A. Y. (Ph.D. in biology) as a science-education specialist who conducted many of these classroom observations, (2) A. F. (Ph.D. in linguistics) as a Project Manager and Research Scientist who investigates sense of belonging in STEM undergraduates and also conducted many of these observations, and (3) R. F. (Ph.D. in chemistry) as a STEM faculty member and a STEM discipline-based education researcher with expertise in qualitative analysis.

During the process, the team determined that two subcategories (1B and 1C) should each be further separated

TABLE II. Categories and subcategories by number of observations for introductory physics courses.

Category	Description	Physics ($N = 42$)
1	Lecture with clicker activities	23 (54.8%)
2	Lecture with demonstrations or other active learning activities	4 (9.5%)
3	Lecture opens with clicker activities (roughly comprising at least one-third of class time)	...
4	Lecture with a combination of clickers, as well as demonstrations or other active learning	15 (35.7%)
Subcategory	Description or criteria	N Physics (42)
1A	No other interaction, limited time spent on clicker	3 (7.1%)
1B-ip*	Limited interaction–Instructor prompted (SQs ≤ 5), presence of clicker activity	6 (14.3%)
1B-sp*	Limited interaction–Student prompted (SQs > 5), presence of clicker activity	2 (4.8%)
1C-ip*	Medium interaction – Instructor prompted (SQs ≤ 5), presence of clicker activity	7 (16.7%)
1C-sp*	Medium interaction–Student prompted (SQs > 5), presence of clicker activity	3 (7.1%)
1D	High interaction, presence of clicker activity	2 (4.8%)
2A	Limited demonstration or other active learning	3 (7.1%)
2B	Frequent demonstration or other active learning	1 (2.4%)
4A	Some interaction, presence of clicker activity, as well as demonstrations or other active learning activities	12 (28.6%)
4B	High interaction, presence of clickers, as well as demonstrations or other active learning activities	3 (7.1%)

*Note: ip = instructor-prompted, sp = student-prompted, SQ = student questions. Category 3 is not represented in our data because it was a unique category that was created by Solomon *et al.* [69] to fit one instructor's teaching style.

based on whether the interaction tended to be instructor prompted (ip) or student prompted (sp). In addition, the team decided category 2 should be split into two subcategories (2A and 2B), based on the total amount of time spent on "other active learning" strategies. No observations fit into category 3, as it was a unique category that was created by Solomon *et al.* [69] to fit one instructor's teaching style. The final categories and subcategories for the physics observations are listed in Table II (detailed descriptions in Supplemental Material [64]), with sample segmentation in Fig. 1. The same analysis for all the disciplines in this overall study are found in the Supplemental Material [64].

To verify that the visual analysis yielded distinct categories and subcategories, we conducted analysis of variance tests (ANOVAs) and Tukey range tests to confirm that these categories and subcategories significantly differed from one another in pedagogically meaningful ways (Supplemental Material [64]). Similar to Solomon *et al.* [69], these analyses confirmed that all categories in our augmented framework significantly differ in the amount of time spent on clicker activities. Category 4 spends the most time on clicker activities and category 2 spends almost no time on them (for all $p < 0.004$). Category 4 spends significantly less time lecturing and significantly more time conducting demonstrations than either category 1 or 2 (for all $p < 0.001$). All the categories differ significantly on the amount of time spent on "other active learning", with category 2 spending the most time and category 1 spending the least (for all $p < 0.05$).

E. Collection of survey data

All students enrolled in these courses ($N = 5987$) in Fall 2018, Spring 2019 and Fall 2019 were invited to participate in this study, which was approved by the university's Institutional review board. Students received some form of extra credit in their laboratory course or lecture course as compensation for completing the surveys. Approximately 92.8% of students ($N = 5557$) consented to participate. Any student who did not consent to participating was removed from the survey dataset. Data from the registrar were obtained for those who consented to participate, including "gender," which refers to the biological sex of the student that is on their birth certificate. We used this as a proxy for the number of men and women in the course, although we recognize that not all students identify with the binary gender they were assigned at birth and this is an imperfect variable to study gender equity.

Students received a packet of surveys at the beginning and end of the semester in each course. Two survey items about overall course inclusivity and self-reported individual verbal participation came from a survey measure developed by the Center for Integrated Research on Cognition, Learning and Education (CIRCLE) to examine academic inclusion in the course. Students were asked to rate "How inclusive do you feel the course [insert course name] was overall?" on a 1-5 scale, with 1 being "not at all" and 5 being "highly inclusive." The self-reported participation question asked "In [insert course name], please estimate the percentage of class periods during which you individually asked or answered a question out loud" with a response slider ranging from 0% to 100%. These two survey items were administered only at the end of each semester. In addition, six survey items measuring students' belonging in the target course [33] were administered at the beginning and end of each semester, though only the end-of-semester survey responses were utilized in this study. The belonging survey included both a "sense of belonging" factor and a

"belonging uncertainty" measure. A previous study at this institution used factor analysis to establish these as two statistically distinct (or separable) scales, with four items loading onto the perceived belonging factor and two loading onto belonging uncertainty [33], and an additional study at a different university found the same factor analysis [32]; the survey used in these previous studies was adapted from Walton and Cohen's belonging survey [31]. The questions were assessed on a six-point scale, and the survey is listed in the Supplemental Material [64]. Only the four items corresponding to the sense of belonging scale were used in this study and the responses were averaged to create a composite belonging score.

F. Quantitative analyses

To compare the gender equity of individual verbal participation across observations from different classroom contexts, we used two approaches. First, we created a "participation score" measure. The participation score is a likelihood ratio defined as follows:

$$\text{Participation Score} = (P_W/C_W)/(P_M/C_M),$$

where P_W is the proportion of women participants, C_W is the proportion of women in the class, P_M is the proportion of men participants, and C_M is the proportion of men in the class. A participation score of one means that participation is gender neutral; i.e., just as many men and women individually participated verbally as expected given the gender composition of the course. A participation score that is greater than 1 means women participated at a higher rate than expected, and a participation score less than 1 means that men participated at a lower rate than expected. We chose to examine likelihood ratios with women as the numerator not only because this follows previous work [37], but also because this allowed us to retain most of the observation data. Specifically, likelihood ratios are undefined when the denominator is zero, and our dataset included far more observations where no women participated in ($n = 14$) than where no men participated ($n = 4$). In addition, because there were no physics observations lacking men participants (i.e., all 4 were other STEM observations), no study 1 data were lost.

In the analyses described below, we use participation scores when examining the observation data in aggregate; e.g., when testing for differences in the average participation score across active-learning categories (1, 2, 4) or across disciplines (physics versus other STEM courses). Participation scores therefore provide an overall view of the gender balance of individual verbal participation. Nonparametric tests (i.e., one-sample Wilcoxon tests, two-sample Mann-Whitney U tests, Kruskal-Wallis tests, and Spearman's correlation tests) were used to examine significance because our likelihood-ratio distributions were non-normal distributions (Shapiro-Wilk test: $W = 0.79$, $p < 0.001$), and rank-biserial correlations were used to determine effect sizes (0.1- small, 0.3-medium, 0.5-large) [72,73]. Participation scores therefore provide an overall view of the gender balance of individual verbal participation. One limitation of the participation scores is their failure to account for the total number of individual participants in each observation, which may exaggerate the presence of gender inequitable participation when the total number of individual participants is low.

Therefore, we also conducted a second set of analyses using the binomial test, which does factor in the total number of individual participants. The binomial test does not provide a high-level view of gender equitable participation; instead, it assesses whether the individual verbal participation in a single observation deviated from the expected distribution given the gender composition of the class. Thus, we implemented the binomial test to statistically evaluate the hypothesis that men might disproportionately participate in STEM classes at the level of individual observations. We used the binomial analyses to complement and corroborate the aggregate analyses that used the participation score measure. Cohen's h was used to determine the effect size of the difference between the two proportions used in each binomial test (i.e., the proportion of men or women participating in a class session versus the proportion of men or women enrolled; 0.2 = small, 0.5 = medium, 0.8 = large) [74].

Although the primary focus of this investigation was individual verbal participation during class, we also examined the gender balance of individual participation in clicker activities using the binomial test. In other words, we tested whether women and men students responded to clicker questions at expected rates relative to the gender composition of the class. Such analyses could only be conducted for those courses that tracked student participation in clicker activities (57/87, 65.5%). It should be noted that clicker participation was used as part of a participation grade in some courses, which could have influenced a student's likelihood to participate. The physics courses we observed required regular participation as part of their final grade (accounting for 3%–5%, depending on the semester), although they did have built-in absences allowed each semester. In the other STEM courses, students in biology courses could receive up to 1% of their final course grade in extra credit for participating in clicker activities and chemistry students received no credit for their clicker participation. None of the calculus courses used clicker activities.

For the student survey data, Pearson correlation tests examined the bivariate relationships among the three late-semester survey outcomes: self-reported individual verbal participation, course-level sense of belonging, and course inclusivity. In addition, two linear mixed-effects regression

models assessed whether students' self-reported participation in a course correlated with their late-semester sense of belonging and inclusivity at the end of that course, and whether those correlations varied by gender. Fixed effects included self-reported participation rate (centered), gender ($F = -0.5$, $M = 0.5$), and the participation x gender interaction; random intercepts for the course (e.g., Introductory Physics I, Fall 2018; Calculus I, Fall 2019) were also included. The dependent variable was either late-semester belonging or course inclusivity.

III. RESULTS

First, we describe the results of the visual analysis and categorization of faculty, then we describe the results for our two research studies.

A. Visual analysis results

Three of the four categories in Solomon *et al.* [69] were found in our data; category 3, which was limited to a single faculty member in the previous study, was not used in our work. Observations in category 1 use clicker activities as their main active-learning component, while those in category 2 use demonstrations or other active-learning strategies (such as small-group discussion and individual problem solving) as their primary active-learning component with minimal or no clicker questions. Observations in category 4 use multiple pedagogies with no one dominant strategy to create an active classroom (clickers were always used, in addition to demonstrations or other active-learning strategies). We further divided the observations into subcategories, differentiating the implementation styles of the faculty based on frequency of active learning and interaction prompt type (i.e., instructor-prompted, denoted ip, versus student-prompted, denoted sp) (see Table II). In contrast to Solomon *et al.* [69], the current study categorized class sessions on the observation level instead of at the faculty level. This current approach means that a single faculty member could have multiple observations that fell into different categories and/or subcategories.

The visual analysis confirmed that physics instructors successfully enacted their course philosophy by using a variety of learning activities. As shown in Table I, the physics observations were predominately in categories 1B-ip, 1C-ip, 4A, and 4B. They almost always incorporate clicker activities in their courses, but a sizable number (15/42, approximately 36%) fall into category 4, which means they spend a similar amount of time on an additional learning activity (typically demonstrations but occasionally it is other active-learning strategies, such as individual or small-group problem solving). When students individually participate in the physics courses, that participation appears to be mostly instructor prompted (i.e., students are answering instructor questions: ip, $N = 13$; sp, $N = 5$).

B. Study 1: Exploring individual participation patterns in introductory physics courses

1. Is individual verbal participation in the introductory physics courses gender balanced?

When individual participation rates are examined in aggregate over all the physics observations, we see evidence of gender inequity in participation. The average participation score in the physics courses is 0.67, meaning men on average participated more than expected based on the course rosters. A one-sample test confirms that the average participation score ($M = 0.67$, SE $= 0.10$) significantly deviates from a neutral participation score of one ($V = 192$, $p < 0.001$, rank biserial $= -0.57$). When physics observations are broken out by category, the most populated category (category 1) has an average participation score significantly different from one, with the effect size demonstrating a large degree of gender inequitable participation ($V = 52$, $p = 0.005$, rank biserial $= -0.62$; Fig. 2).

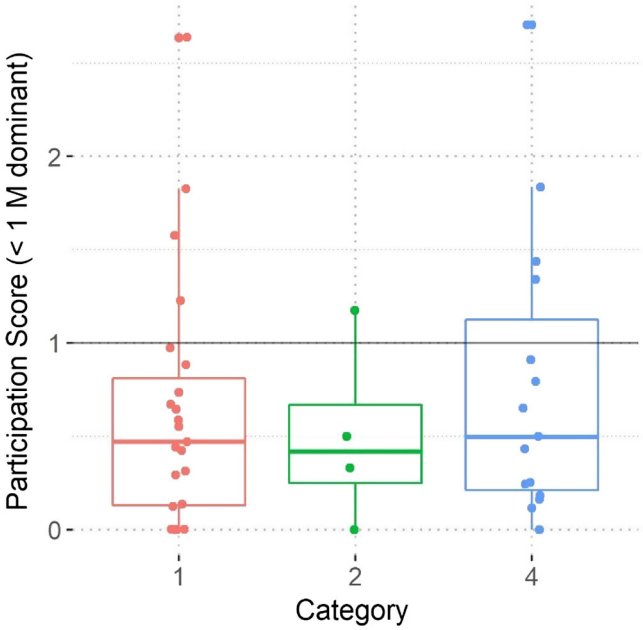

FIG. 2. Examining physics observations by category and participation score. Box-and-whisker plots of the participation scores in all categories within physics. Each point represents a class observation relative to the "neutral" participation score of one (horizontal line). The boxes show the median participation score in each category (center bar), the first and third quartiles (i.e., the 25th and 75th percentiles; "hinges" or outer lines of box), and 1.5 times the interquartile range ("whiskers" or lines extending from boxes). The majority of observations have higher rates of participation of men (i.e., participation scores lower than 1), and the mean participation scores do not differ from each other. Category 1 is for lectures with clicker activities as its main active-learning component; Category 2 is for lectures with either demonstrations or other active learning as its main active-learning component; and Category 4 is a combination of clickers as well as demonstrations or other active-learning components.

Category 2 ($M = 0.50$, SE $= 0.25$) has a smaller mean participation score than category 1 ($M = 0.63$, SE $= 0.14$), suggesting a stronger bias towards men's participation, but it fails to significantly differ from the neutral score of one ($V = 1$, $p = 0.10$, rank biserial $= -0.80$), probably due to the small number of observations in that category ($n = 4$). The mean participation score for category 4 is closer to one ($M = 0.77$, SE $= 0.20$), but the difference approaches significance with a medium effect size ($V = 34$, $p = 0.07$, rank biserial $= -0.43$). The categories do not differ from each other in terms of mean participation score ($\chi^2 = 0.45$, $p = 0.7974$).

As mentioned in Sec. II, the number of individual participants in an observation is not accounted for when calculating a participation score, potentially inflating the impact of observations with few individual participants. Therefore, we ran binomial tests to assess gender equity in participation at the individual observation level and calculated Cohen's h (0.2 = small, 0.5 = medium, 0.8 = large) [74] to determine the direction and magnitude of the difference between the two proportions (i.e., the proportion of participants that are men and proportion of men on the roster), while adjusting for the total number of participants in the target class session. Thirty-three (78.5%) of the physics observations had a negative Cohen's h, indicating more men participated than expected based on the roster (Fig. 3). For 10 of those 33 observations, the descriptive gender participation imbalance observed using Cohen's h reached statistical significance ($p < 0.05$). Thus, almost a quarter (23.8%) of individual physics observations ($N = 42$) were not gender equitable in participation, with men speaking up in class at disproportionately high rates.

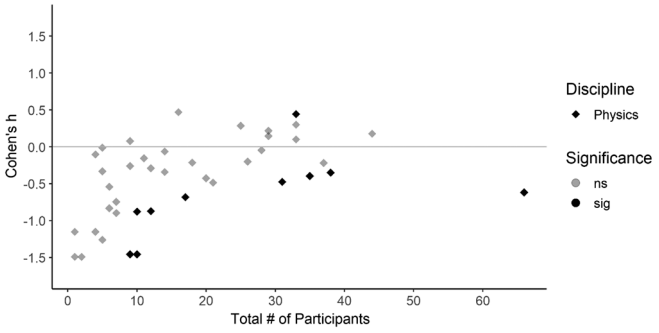

FIG. 3. Examining individual classroom observations for gender equity. Each point represents a binomial test evaluating whether the observed proportion of men individually participating verbally in class differs from the proportion of men on the roster. Cohen's h represents effect size, or the distance between the expected proportion (based on the course roster) and the observed proportion of men participating. A Cohen's h greater than zero indicates that women participated at higher levels than expected, while a Cohen's h less than zero indicates that men participated at higher levels than expected. Significance is denoted by color: the 10 (out of 42) significant tests are black, and the nonsignificant tests are gray.

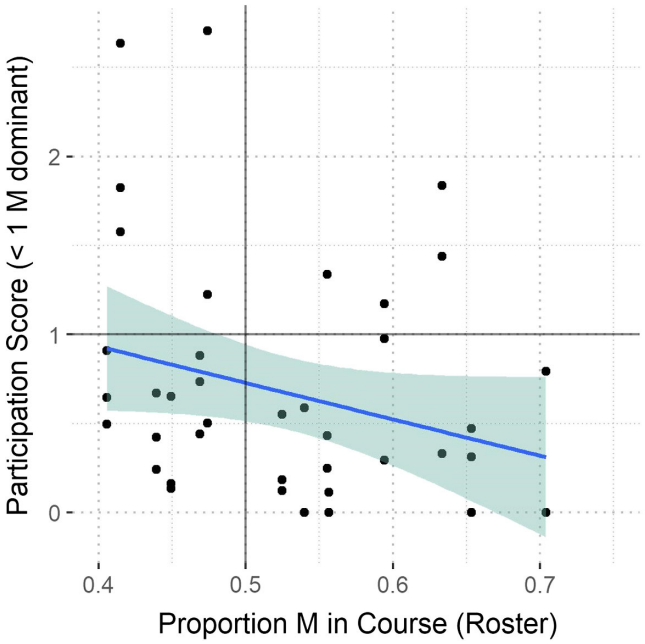

FIG. 4. . Women tend to individually participate more in women-majority physics classes. Scatterplot of physics observations by gender representation in the course (x axis) and participation score (y axis). Simple regression lines illustrate the linear relationship between proportion of men in the course and participation scores (shaded area represents the 95% confidence interval). We find a moderate, significant correlation between participation scores and the proportion of men and women enrolled in the course, meaning we tend to find higher rates of women individually participating when there are more women in the classroom.

Although both the participation scores and the binomial tests account for the gender composition of the course, we explored whether the pattern of gender in participation in physics was linked, at least in part, to representational inequities in the student population. A moderate, significant correlation emerged between participation scores and the proportion of men and women enrolled in the course ($\rho = -0.31$, $p = 0.05$) (Fig. 4). In other words, we found an association between our aggregate measure of participation and gender representation in the roster.[1] This relationship has also been seen in Bailey *et al.* [27], where females participated more in biology classes when they were a student majority. Taking all of these pieces together, the observational evidence therefore suggests that gender inequity in individual verbal participation in introductory physics may be linked to course enrollment patterns.

[1]Where available, clicker data, which show the number and gender of students in the classroom during a specific observation, confirm the roster provides an accurate description of gender representation in classroom [$r(35) = 0.939$, $p < 2.2 \times 10^{-16}$].

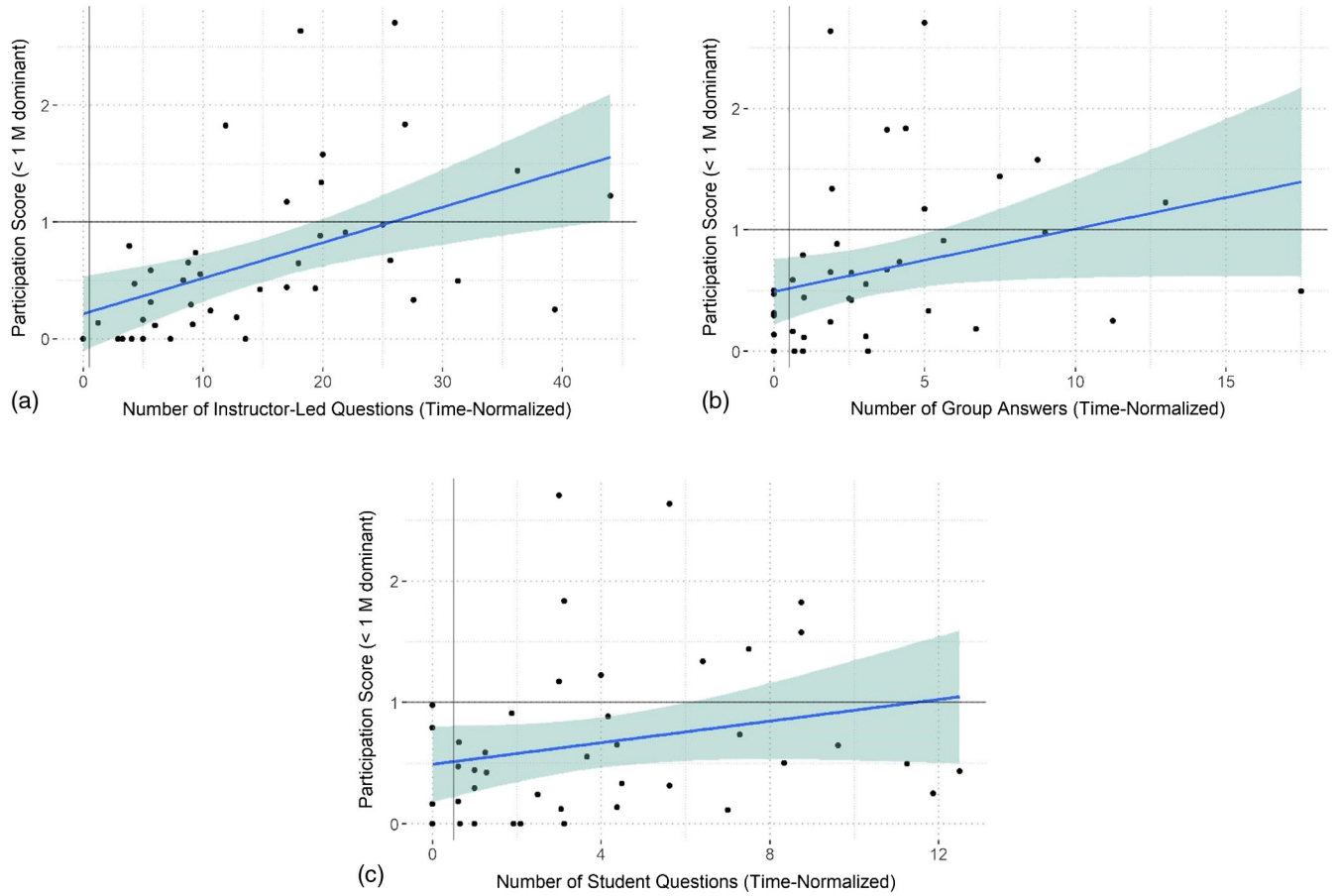

FIG. 5. An increase in the number of (a) instructor-prompted questions, (b) group responses, and (c) student questions appears to be strongly correlated with a higher participation score (regardless of category). Scatterplots of physics observations by the (time-normalized) number of instructor questions (x axis) and participation score (y axis). Simple regression lines illustrate the linear relationship between instructor-led questions and participation scores (shaded area represents the 95% confidence interval).

2. Are more interactive classes correlated with more equitable participation? Are there any instructor or student behaviors that correlate with equal participation?

While many observations had participation scores less than 1, others had a near neutral participation score or scores greater than 1. We therefore sought to determine if any instructor or student behaviors correspond with more equitable individual participation. We examined the relationships between participation scores and the number of (a) questions an instructor asked, (b) group responses by students in the course, (c) questions students asked, and (d) clicker questions asked per class period.

Within physics classes, we found a significant, large correlation between the number of instructor questions and participation scores ($\rho = 0.63$, $p < 0.001$). As seen in Fig. 5(a), more instructor-prompted questions corresponded with higher participation scores, indicating individual participation more closely reflected the gender makeup of the class. This finding appeared across instructors and categories. We also found a significant, moderate correlation between the number of group responses by students and participation scores within physics classes ($\rho = 0.54$, $p < 0.001$) [Fig. 5)], although the interpretation of this pattern remains unclear. More active classes (i.e., the classes with higher numbers of individual participants) tended to have higher numbers of group answers [$r(40) = 0.60$; $p < 0.001$], suggesting that the association between equitable participation and group responses could reflect an underlying tendency for more active class sessions to be more equitable. The number of student questions was also found to significantly correlate with more equitable participation [$\rho = 0.33$, $p = 0.04$; Fig. 5)]. This potentially signals that if students are invited to ask questions, women may be more likely to participate. The number of clicker questions did not significantly correlate with more equitable participation ($\rho = -0.10$, $p = 0.54$).

3. Is individual participation via clickers equitable?

To examine whether individual participation via clickers was gender equitable, we ran binomial tests comparing the proportion of men and women who submitted clicker responses with the proportion of men and women enrolled in each class section. The tests proved nonsignificant for all classroom observations that used clicker activities ($N = 37$,

TABLE III. Summary of participants' self-reported individual verbal participation in each semester.

	Fall 2018			Spring 2019			Fall 2019		
Department	*F* % (n; SE)	*M* % (n; SE)	*t*	*F* % (n; SE)	*M* % (n; SE)	*t*	*F* % (n; SE)	*M* % (n; SE)	*t*
Physics	9.60 (269; 1.17)	15.13 (316; 1.19)	−3.30*	9.60 (215; 1.23)	14.44 (255; 1.43)	−2.57*	11.22 (255; 1.35)	15.91 (240; 1.52)	−2.31*

Note: Percentages indicate the proportion of class sessions that students perceived themselves individually and verbally participating. Subsample sizes broken out by gender are reported in parentheses (n). Two-sample t tests are reported, and * indicates $p < 0.05$.

$p > 0.10$). Therefore, in contrast to the individual verbal participation data, clicker participation rates did not favor men or women (i.e., men and women equally participated in clicker activities, given their representation in class).

4. Do student perceptions of their individual verbal participation in the introductory physics class mirror the observational data?

To address this question, we averaged the self-reported individual verbal participation of men versus women in each overall course (i.e., combining all the survey participants across all the physics sections in a given semester). We found that men consistently reported higher participation rates than women ($p < 0.023$, see Table III); on average, women reported participation in 10% of all class sessions and men reported participation in 15% of all class sessions. This pattern aligns with the classroom observation data, which illustrated higher levels of individual verbal participation by men.

5. Does self-reported individual verbal participation correlate with course-level belonging and inclusivity at the end of the course?

Both course-level sense of belonging and inclusivity at the end of the semester showed small, significant correlations with self-reported participation levels [belonging: $r(1547) = 0.19$; inclusivity: $r(1547) = 0.16$, $p < 0.001$]. The positive correlations indicate that greater belonging and inclusivity are associated with more frequent participation (i.e., self-reported participation in a larger percentage of class sessions). Belonging and inclusivity scores were also found to be moderately correlated with one another [$r(1547) = 0.49$, $p < 0.001$], which is not surprising given that classroom inclusivity affects course-level sense of belonging [35].

Linear regressions were used to explore whether course-level sense of belonging and inclusivity, as dependent variables, relate differently to self-reported participation among men versus women. As shown in Table IV, neither regression revealed a significant interaction of gender and participation ($p > 0.29$). Instead, all students who reported participating more often tended to report stronger feelings of belonging and inclusivity, regardless of gender. In addition, a main effect of gender indicates an overall tendency for men to report stronger belonging than women ($p < 0.001$), as seen in other STEM studies on belonging [32,33]; gender was not associated with inclusivity.

C. Study 2: Comparing individual participation patterns in physics to other introductory STEM courses

In study 2, we compared the findings of the introductory physics class sessions to other introductory STEM classes (specifically introductory biology, general chemistry, and calculus) as a collective group, which had different amounts and patterns of active learning, class sizes, and proportions of men and women. In addition, we checked whether the correlation patterns observed in introductory physics are seen in these other disciplines. For instance, are instructor-prompted questions also associated with more equitable individual participation in other introductory STEM courses? More detailed information about the methods and results for the other STEM disciplines can be found in the Supplemental Material [64].

1. Similarities between physics and other STEM

Many of the patterns seen in the study 1 observational data were also replicated in study 2. Like physics, the other

TABLE IV. Linear mixed effects regression models predicting physics post-survey outcomes.

A. Sense of belonging model				
Predictor	B	SE	*t*	*p*
Intercept	4.61	0.05	94.05	<0.001*
Participation	0.007	0.001	6.98	<0.001*
Gender	0.19	0.04	4.35	<0.001*
Participation x Gender	0.001	0.002	0.55	0.58
B. Inclusivity model				
Predictor	B	SE	*t*	*p*
Intercept	4.11	0.08	51.09	<0.001*
Participation	0.007	0.001	6.12	<0.001*
Gender	0.01	0.05	0.23	0.81
Participation x Gender	0.002	0.002	1.06	0.29

Note: $N = 1549$. * denotes $p < 0.05$. Self-reported participation was centered, and gender was contrast-coded ($F = -0.5$, $M = 0.5$). Both models include random intercepts for course ($n = 3$; Physics 1 in Fall 2018 and 2019, and Physics 2 in Spring 2019). Sense of belonging was measured on a 6 pt. scale, and inclusivity on a 5 pt. scale.

STEM disciplines had an average participation score that differed from one and were less than 1 ($V = 247$, $p = 0.009$, rank biserial=-0.43), indicating that men were overrepresented among individual verbal participation. The mean participation score in the other STEM courses ($M = 0.85$, SE $= 0.16$) did not significantly deviate from the physics' mean ($M = 0.67$, SE $= 0.10$; $W = 939.5$, $p = 0.48$), suggesting similar levels of gender inequitable participation across STEM disciplines (Fig. 6). Like physics, the majority of binomial tests in other STEM courses (66.7%) also showed participation favoring men (indexed by negative Cohen's h scores, Supplemental Material [64]), yet the number of observations with statistically significant binomial tests was smaller in other STEM courses ($N = 4/45$; 8.8%) compared to physics ($N = 10/42$; 23.8%).

In terms of instructor and student behaviors, a significant, moderate, positive correlation was once again observed between the number of student questions and participation scores in the other STEM courses ($\rho = 0.31$; $p = 0.05$), parallel to physics. Also comparable to physics,

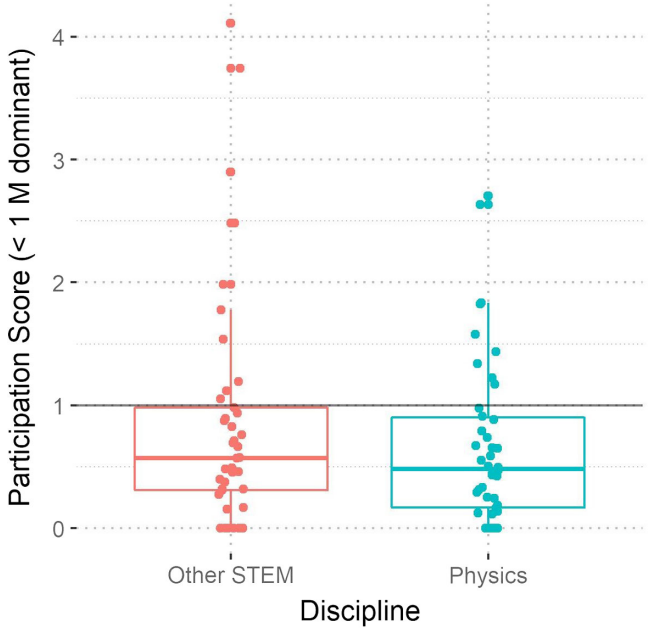

FIG. 6. Participation scores across all disciplines seem to show a disproportionate amount of men individually participating. Box-and-whisker plots of the participation scores in physics courses versus all other introductory STEM courses observed. Each point represents a class observation relative to the "neutral" participation score of one (horizontal line). The boxes show the median participation score in each category (center bar); the first and third quartiles (i.e., the 25th and 75th percentiles; "hinges" or outer lines of box), and 1.5 times the interquartile range ("whiskers" or lines extending from boxes). While the other STEM courses had a higher average participation score ($M = 0.85$) than physics ($M = 0.67$), their means did not significantly deviate from each other ($W = 939.5$, $p = 0.48$), suggesting similar levels of gender inequity in individual participation across disciplines.

there was no significant correlation between the number of clicker questions and other STEM participation scores ($\rho = 0.17$, $p = 0.29$).

In all the other STEM classroom observations that used clicker activities ($N = 21/45$), binomial tests once again showed no significant differences between the gender balance of clicker participation and the gender representation in the course rosters (Supplemental Material [64]). The consistency of this result across physics and the other STEM courses reinforces the idea that clicker activities elicit equitable participation in the classroom. At the same time, the lack of correlation between the number of clicker questions and participation scores (see above) suggests that equitable participation via clickers does not translate into more individual women participating verbally.

Similar to physics, the student self-participation data in the other STEM courses aligned well with the observational data. As mentioned above, 2/3 of the binomial tests for the other STEM courses generated negative Cohen's h scores, which indicate class sessions where participation was biased towards men. However, the binomial tests reached significance for only four class sessions (1 math, 3 biology; Supplemental Material [64]), suggesting that the gender inequitable participation was not very reliable across the other STEM observations. Paralleling these observation data, the survey data show that men tend to self-report descriptively higher levels of participation than women in most other introductory STEM courses, but the gender difference in these self-reports only reaches significance in a single course, the Fall 2019 semester of Introductory Biology [$t(268.8) = -2.12$, $p = 0.04$; Supplemental Material [64]]. This is also the same course where the binomial tests for 2 classroom observations reached significance.

If we directly compare students' self-reported participation levels across disciplines, regression analysis confirms that both physics and the other STEM courses exhibited an overall effect of gender, with men in all disciplines reporting 3.35% higher individual verbal participation than their women classmates [$B = 3.35$, SE $= 0.61$, $t(3822.27) = 5.47$, $p < 0.001$]. However, a significant interaction between gender and discipline in the self-reported participation rates was larger in physics ($M > F$ by 5.04%, $t(3820.9) = -5.36$, $p < 0.001$) than in the other STEM courses [$M > F$ by 1.66%, $t(3825.2) = -2.12$, $p = 0.15$]. These results suggest that the gender inequity of individual verbal participation was less robust in the other STEM courses than in physics.

Finally, both sense of belonging and inclusivity at the end of the semester in the other STEM courses showed significant correlations with self-reported participation levels [belonging: $r(2185) = 0.11$; inclusivity: $r(2279) = 0.09$, $p < 0.001$]. The significance of these correlations matches the physics results, but they are weaker in strength, barely meeting the threshold for a small effect. As in physics, sense of belonging and inclusivity scores were moderately

correlated with one another in the other STEM courses [$r(2224) = 0.39$, $p < 0.001$].

2. Differences between physics and other STEM

While many of the findings in physics (i.e., study 1) were replicated in the other STEM courses (i.e., study 2), a few differences are found between physics and other STEM disciplines. In physics, a moderate significant correlation between an increasing participation score and a decreasing proportion of men in the classroom ($\rho = -0.31$, $p = 0.05$; Fig. 4), meaning that representation in the classroom may have an effect on women's individual participation. This finding was not replicated in the other STEM courses, where a small, but nonsignificant, correlation in the opposite direction was found ($\rho = 0.23$, $p = 0.15$). Therefore, the evidence of inequitable participation cannot be explained by gender representation in the classroom in other STEM classes, similar to other previous work [24,25]. In physics, we saw a strong association between an increasing number of instructor questions and an increasing participation score ($\rho = 0.63$; $p < 0.001$); in the other STEM courses, this correlation is not significant and is in the opposite direction ($\rho = -0.25$, $p = 0.11$; Supplemental Material [64]). One potential explanation for these contrasting results is variation in class size; the observed physics classes tended to fall around 100 students in size, whereas some of the other STEM classes were as large as 322 students. When we filtered out the other STEM classes larger than 150 students (the gender participation maps showed that actual attendance of these courses is around 100 students, which is comparable to physics courses), the correlation switched to the same direction as physics but remained nonsignificant ($\rho = 0.09$, $p = 0.75$; Supplemental Material [64]). This was also found with group responses as well, where the strong correlation seen in physics disappeared and instead found a small but not significant correlation in the opposite direction ($\rho = -0.28$, $p = 0.08$; Supplemental Material [64]). We again looked at the subset of other STEM classes having a similar size as physics, and this time the correlation did not switch to match physics, and there was still no significance ($\rho = -0.14$, $p = 0.61$).

IV. DISCUSSION

Although the focus of this investigation (gender equity in individual participation) differed from Solomon *et al.* [69] (variety in clicker implementation), the general categorization structure still applied to our new set of observations [69]. The same basic categories differentiated sessions where active learning consisted primarily of clicker activities (category 1), demonstrations or "other active learning" (category 2), or a mix of at least two different pedagogies (i.e., clickers are used, in addition to demonstrations or other active-learning strategies; category 4). At the same time, we expanded this category structure to accommodate the increasingly complex ways faculty incorporate active learning and interaction into their introductory STEM courses. We refined the initial framework by further dividing certain subcategories (1B and 1C) to reflect how different instructors rely more heavily on either instructor-prompted or student-prompted interaction. We also separated category 2 into subcategories 2A and 2B, acknowledging variety in the amount of time instructors chose to spend on other active-learning techniques (e.g., individual problem solving, or small-group discussion). Through these changes, we developed a modified framework consistent with previous research [69] that also captured the nuance of our faculty's evolving use of active-learning strategies.

A. Use of active-learning strategies is not synonymous with equity

Active-learning courses are often promoted for their potential to increase equity in STEM classrooms. However, like Aguillon *et al.* [25] concluded about an introductory biology course, this investigation suggests that active learning is necessary but not sufficient to induce gender equitable participation in introductory STEM courses across several semesters, at least in terms of individual verbal participation. The physics courses we observed had an average participation score less than 1, indicating disproportionately higher individual verbal participation by men and lower individual verbal participation by women relative to their proportion in the class. Observation-level analysis confirmed that three-quarters (76.2%) of the physics observations had descriptively higher than expected participation by men and almost a quarter of observations (23.8%) were statistically significant for a gender imbalance in individual verbal participation. A similar pattern emerged in the other STEM courses, where the average participation score once again indicated disproportionately high participation by men. However, observation-level analysis in the other STEM courses showed this gender imbalance still to be present but less pervasive, with 66.7% of the other STEM observations demonstrating the descriptive pattern of participation biased towards men and only 8.9% of observations statistically significant. Importantly, a gender imbalance in participation still emerges in classrooms in which women are a majority; therefore, the problem is not merely solved by getting more women into STEM or the classroom. Proactive strategies are needed to create equitable individual verbal participation in the classroom.

Within physics, we found that three classroom behaviors correlated with a higher participation score (i.e., a score closer to 1.0, reflecting more gender equitable participation): the number of questions posed by an instructor, the number of group responses by students, and the number of student questions. However, only one of these practices was correlated with an increasing participation score in the other STEM courses (number of student questions).

Suspecting that variety in class size in the other STEM courses could mask the effects of different classroom behaviors, we examined a subset of classes similar in size to physics (<150 students) and found the correlations resembled those in physics courses, however they still were not significant (possibly due to lack of power; $n = 16$). We suspect that larger introductory STEM classes might feel overwhelming to students, so pedagogical techniques used to encourage participation in medium-sized classes (80–150 students) may not quite have the same effect in larger courses (>150 students). While we hypothesize that class size may contribute to the differential impact of these pedagogical techniques on equitable participation in physics versus the other STEM courses, the current sample cannot systematically test this theory.

Despite this limitation, the current investigation shines a light on the complex dynamics of active learning in STEM classrooms. For example, while group answers may be a valuable practice, and they correlate with more gender equitable participation in our physics observations, the technique seems unlikely to be the underlying reason behind more equitable individual verbal participation. More frequent group answers also corresponded with a higher number of individual participants, suggesting that at least within physics, more frequent group answers are an indicator of class sessions that are generally more active and perhaps more equitable as a result. At the same time, this study also suggests that more active-learning opportunities does not automatically translate into more gender equity in individual verbal participation. Physics classes in our sample spent less time lecturing than the other STEM courses ($p = 0.03$), but the gender disparities in physics participation proved larger and more pervasive than in the other STEM courses. The combined findings of study 1 and study 2 provide evidence that the adoption of active-learning strategies may not foster equitable participation in and of itself. Instead, proactive steps need to be taken to facilitate participation during those activities so students from all identity groups are equally likely to respond.

B. Inhibitory influences on women's participation

Why are women less likely to individually participate out-loud in class? We hypothesize several factors might be at play. First, women may already feel they have participated in other ways. As mentioned in the introduction, Fritschner [39] observed that students can have varying definitions of "participating" in a course. These different perceptions could be due to the way a course outlines participation points in the syllabus (e.g., all the biology and physics courses we observed provide incentives to "participate" in class via clickers). Other women may feel that answering questions via a group response is sufficient participation and a way for them to engage with the material without singling themselves out. Especially if women already have participated through other strategies, they may feel less desire to volunteer an individual response due to higher levels of anxiety about individual verbal participation compared to men [28]. However, further work examining students' views of participation is needed, especially in larger class sizes.

Another factor that might affect women's individual verbal participation levels is stereotype threat [26], or fear of confirming negative stereotypes about women's scientific and quantitative abilities. Previous work has shown that men are more likely to be named by their peers as knowledgeable about course content in an introductory biology course, and this bias persists even when controlling for class performance [75]. Another study identified that women were more likely to describe feeling anxious when asking questions, or feel they don't know the material well enough to ask questions in their large enrollment science courses [29]. In addition, we noted in the introduction that men are more likely than women to perceive themselves as a "physics person" [14,15], and report a higher sense of belonging in physics [13,14]. This finding is confirmed in our study, where women consistently report lower averages of belonging in physics courses, and prior research links lower physics belonging among women to negative gender stereotypes [13]. Moreover, recent work has established that women students in STEM fields, including physics, experience higher levels of belonging uncertainty than men, and such uncertainty might undermine women's motivation to participate [32]. Therefore, women may be less likely to speak out in class because of their fear of appearing or being perceived as less capable than their classmates.

In contrast to individual verbal participation, our findings suggest that individual participation via clicker activities is equitable in all the STEM courses we observed. The proportion of men who participate via clicker is strongly correlated with the proportion of men enrolled in the class section [$r(55) = 0.92$, $p < 0.001$]. In other words, there is a close correspondence between gender representation in the classroom and clicker participation, with men and women answering clicker questions at expected rates. Clicker questions provide a way for students to engage with the material, but their answers are anonymous to the rest of the class (unless they choose to share). Gender-balanced participation with clickers potentially points to anonymous responding as a path towards more equitable participation. However, verbal versus clicker participation may not always be equivalent in terms of student learning (i.e., both may not require the same level of cognitive engagement) [76].

C. Student perceptions mirror observations

Our classroom observation data aligned well with students' self-reported participation data, demonstrating the value of a multimethod research study for cross-validating results. Although some literature has previously

shown that self-reported participation data from students are not always accurate [49,77,78], our student-generated data largely preserve the patterns of individual verbal participation we observed among men and women in class. Through observation data, we found a consistent tendency for men to individually participate at higher levels than women, creating a significant overall gender effect, which was driven particularly by physics courses. This pattern was echoed in the self-reported participation data, where on average men described higher levels of individual verbal participation than women in every single discipline (Table III), although these gender inequities only reached statistical significance in a subset of semesters (all semesters of physics and F19 semester of Introductory Biology). Nonetheless, the finding of less robust gender inequities in the self-reported participation data from the other introductory STEM courses compared to physics is notably similar to the significance patterns in the observation-level binomial tests.

Although the correlations between self-reported participation and course-level sense of belonging or inclusivity cannot speak to the causal nature of their relationship, the results illustrate a link between students' classroom behaviors and their subjective experience of STEM courses. We found that stronger feelings of belonging and inclusivity in the course were positively correlated with higher levels of self-reported participation. We hypothesize that creating ample, equitable opportunities for students to (verbally) participate individually may foster more universal feelings of belonging and inclusivity in the course [35]. Conversely, students who feel a stronger sense of belonging and inclusivity may feel more comfortable taking advantage of opportunities to verbally participate individually. Either way, our study shows a connection between students' behavioral and affective engagement in STEM courses, and it emphasizes the importance of facilitating active-learning activities in a manner conducive to creating equitable verbal participation.

V. LIMITATIONS

This study is subject to some limitations which may affect the interpretation of its results. One limitation is that the vast majority of our clicker data is associated with some small point value in the course, which makes it difficult to evaluate if clickers do indeed lead to more equitable individual participation. However, we find the same trend of gender equity in individual participation in all our clicker data, which include courses that give points towards the course grade, extra credit, or no credit at all for participating in the clicker activities. In fact, among observations in which students received either extra credit or no credit for participating, where participation might be considered more of a choice than an obligation, 86% (18/21) of the observed classes have women participating via clickers more than expected based on the course roster. Although the trends in our data do look promising, further work is needed before reaching a conclusion about the causal relationship between clicker activities and gender equity in participation.

Another limitation of our study is the confound between class size and discipline, which makes it difficult to isolate their independent effects. The majority of introductory courses within each discipline at our institution tend to be the same size, but Fall 2019 observations provided the opportunity to investigate cross-discipline variation in the gender equity of verbal participation. We examined a subset of other introductory STEM course observations that matched the physics observations in terms of class size, to compare disciplines within a sample of smaller-sized classes. Although the associations between participation score and classroom behaviors in the matched other STEM sample looked descriptively similar to physics, they generally failed to reach significance. Such results may reflect a lack of power, or they may indicate that we collapsed across variation in the other STEM disciplines. Either way, further work is needed to rule out meaningful differences between physics and other STEM disciplines, and to explore the role that class size, among other factors (e.g., disciplinary culture, instructor attitudes), plays in individual participation equity.

A third limitation of this study stems from examining gender equity in verbal participation without distinguishing different forms of participation (e.g., questions versus answers) or the timing of participation (e.g., in follow-up discussion versus during lecture). Unlike the study of Eddy *et al.* [24], which isolated gender inequity to a specific form of individual verbal participation—namely, voluntarily answering an instructor's question—this study tested for an overall pattern of gender inequity across all instances of individual verbal participation and explored its relationship to students' experience of their classes. In future work, we may want to reexamine individual verbal participation by participation type or timing to test specific hypotheses about the drivers of gender inequity. For example, different theories might predict more equitable participation when examining student questions during the lecture portion of courses, or more pronounced inequities in participation in the follow-up discussion of activities. By investigating gender equitable participation in a more precise way, we may be able to develop more targeted strategies to combat potential inequities.

A fourth limitation mentioned in Sec. II is that the observers inferred the gender of the student in the observations, and therefore, it is an imperfect variable to study gender disparities in the classroom. In addition, treating gender as a binary variable fails to acknowledge the range of different gender identities that students may express. Unfortunately, having students self-describe their gender was not feasible in these large classes, especially because there was no assigned seating, and in many cases, students could attend different sections of the same course. However, all the data we

collected (course rosters, clicker data, self-reported participation data) aligns with our observational data, suggesting that perceived gender is a reasonable proxy variable.

A final limitation which was also noted in Sec. II, is that the gender participation map does not keep track of individuals who participate multiple times in a session. Therefore, the tool may overestimate the number of students who are individually participating, and it may allow highly vocal individual participants to exert greater influence on the gender equitable participation in the observations (e.g., having one or two men repeatedly volunteer answers or questions). While the data may not capture students who individually participate multiple times in a session, it still accurately reflects how often students hear men or women voices speak in the course. In other words, the data do reflect students' potential experience of class sessions in that the gender of individual verbal participants fails to represent the gender composition of the class. It is also important to note that the observational data mirrors the student self-reported levels of participation, with both methods confirming that men are more likely to individually participate than women in their introductory STEM courses.

VI. IMPLICATIONS FOR INSTRUCTORS

This study strengthens the argument that the mere incorporation of active learning does not automatically lead to gender equitable verbal participation in the classroom [24,25,37]. Instead, proactive strategies must be used to ensure that diverse students can and do seize the opportunity to individually participate aloud in class. This section highlights several suggestions for facilitating equitable participation, further described in these resources [79,80].

Instructors can begin by reflecting on the ways they encourage participation in the classroom, making sure their expectations for participation in the course are outlined or defined in the syllabus. Instructors should also verbally communicate their expectations to the students starting on the first day of class, and continuing throughout the semester. In addition, they can motivate students to meet those expectations, explaining how participation can enhance learning and supports course objectives. By being transparent with their students, instructors can help students recognize, appreciate, and fulfill the expectations for the course.

As discussed in Tanner [79], to foster a safe environment that encourages all students to verbally participate, instructors can normalize mistakes and emphasize the collaborative, community aspects of learning. Instructors can choose their words carefully when a student responds with an incorrect answer, working through the student's thought process and highlighting correct steps in the process (e.g. "I see how you came to that conclusion. Let's think about this assumption in the problem; how might that change your conclusion?"). Also, instructors can model behaviors they expect from students. If they make an error in class, they can treat it as an opportunity to affirm that everyone makes mistakes. This approach will create an environment where students may be more likely to participate because they will feel less pressured to provide the correct answer and instead contribute to the learning process.

Another way to encourage participation is to create many, accessible participation opportunities. Instructors may prevent the same few students from always answering questions posed to the class by, for example, resisting the urge to always choose the first hand that goes up, even if doing so makes it faster to progress through the day's material. Allowing a small number of students to dominate participation can create a "consolidation of responsibility" [49], where students who do not regularly participate will become passive participants by relying on those students that regularly participate. An instructor might continuously challenge themselves to ask for volunteers from different parts of the room, or for new voices. These actions can help encourage new students to verbally participate, and instill a sense of shared responsibility for answering questions.

In this study, clicker activities appeared to be a more equitable form of individual participation, suggesting that anonymous individual response strategies may induce more balanced participation. By allowing anonymous responses, clicker activities encourage even the most hesitant students to engage with the course material. They also help students check their understanding and detect misconceptions they may have, in addition to seeing how well they understand the concepts compared to their peers [81]. While there are many advantages to using clicker activities, we found that gender-balanced clicker participation does not necessarily translate into equitable individual verbal participation. Therefore, instructors may need to explicitly encourage students to find their voices in class, which may ultimately help them advance into their field of interest.

Finally, as we asserted at the beginning of the discussion, we cannot assume that gender equity in individual verbal participation is a natural consequence of getting more women in the room. In this study, we still found more men participating than expected even in women-majority classrooms: 63% (55/87) of our observations were women-majority in terms of enrollment, while only 32% (28/87) had more women individually verbal participate than men. While many universities continue the important work of getting more women into STEM classrooms, more effort needs to be channeled into classroom-level strategies that may help women take full advantage of learning opportunities, experience a sense of belonging and inclusivity in the field, and ultimately advance on par with men in their careers.

ACKNOWLEDGMENTS

This research was supported by an Inclusive Excellence grant from the Howard Hughes Medical Institute. We thank

the instructors involved in this study for allowing their courses to be observed. We would also like to thank our observers for their help collecting the data: Chris Wally, Elise Walck-Shannon, Siera Stoen, JD Young, and Jia Luo. Finally, we would like to thank Erin Solomon and Shaina Rowell for assisting in the initial development and piloting of the gender participation map. A. Y. and R. F. conceptualized and developed the gender participation map. A. F. and E. W. S. piloted the gender participation map. A. Y. revised the map with input from the observation team. A. Y., A. F., S. S., C. W., E. W. S., J. L., and J. Y. collected the observational data. A. Y., A. F., and R. F. conducted the visual analyses, developed the paper and wrote the manuscript. A. Y. and A. F. also conducted the statistical analyses.

[1] National Science Foundation *Higher Education in Science and Engineering* (National Science Foundation, Alexandria, VA, 2019), https://ncses.nsf.gov/pubs/nsb20197/executive-summary (accessed January 13, 2020).
[2] J. Clark Blickenstaff, Women and science careers: Leaky pipeline or gender filter?, Gender Educ. **17**, 369 (2005).
[3] A. M. Porter and R. Ivie, *Women in Physics and Astronomy 2019* (AIP Statistical Research Center, College Park, MD, 2019), www.aip.org/statistics (accessed January 13, 2020).
[4] National Science Foundation, *National Center for Science and Engineering Statistics, Women, Minorities, and Persons with Disabilities in Science and Engineering: 2017* (Arlington, VA, 2017).
[5] X. Chen, *STEM Attrition: College Students' Paths Into and Out of STEM Fields Statistical Analysis Report* (National Center for Education Statistics, Washington, D.C., 2013).
[6] S. M. Turnbull, K. Locke, F. Vanholsbeeck, and D. R. J. O'Neale, Bourdieu, networks, and movements: Using the concepts of habitus, field and capital to understand a network analysis of gender differences in undergraduate physics, PLoS One **14**, e0222357 (2019).
[7] M. M. Handelsman, W. L. Briggs, N. Sullivan, and A. Towler, A measure of college student course engagement, J. Educ. Res. **98**, 184 (2005).
[8] A. M. L. Cavallo, W. H. Potter, and M. Rozman, Gender differences in learning constructs, shifts in learning constructs, and their relationship to course achievement in a structured inquiry, yearlong college physics course for life science majors, School Sci. Math. **104**, 288 (2004).
[9] C. Lindstrom and M. D. Sharma, Self-efficacy of first year university physics students: Do gender and prior formal instruction in physics matter?, Int. J. Innov. Sci. Math. Educ. **19**, 1 (2011).
[10] E. Marshman, Z. Y. Kalender, C. Schunn, T. Nokes-Malach, and C. Singh, A longitudinal analysis of students' motivational characteristics in introductory physics courses: Gender differences, Can. J. Phys. **96**, 391 (2018).
[11] J. M. Nissen and J. T. Shemwell, Gender, experience, and self-efficacy in introductory physics, Phys. Rev. Phys. Educ. Res. **12**, 020105 (2016).
[12] K. A. Shaw, The development of a physics self-efficacy instrument for use in the introductory classroom, AIP Conf. Proc. **720**, 137 (2004).
[13] J. G. Stout, T. A. Ito, N. D. Finkelstein, and S. J. Pollock, How a gender gap in belonging contributes to the gender gap in physics participation, AIP Conf. Proc. **1513**, 402 (2013).
[14] V. Seyranian, A. Madva, N. Duong, N. Abramzon, Y. Tibbetts, and J. M. Harackiewicz, The longitudinal effects of STEM identity and gender on flourishing and achievement in college physics, Int. J. STEM Educ. **5**, 40 (2018).
[15] Z. Hazari, P. M. Sadler, and G. Sonnert, The science identity of college students: Exploring the intersection of gender, race, and ethnicity, J. Coll. Sci. Teach. **42**, 82 (2013).
[16] S. Lauer, J. Momsen, E. Offerdahl, M. Kryjevskaia, W. Christensen, and L. Montplaisir, Stereotyped: Investigating gender in introductory science courses, CBE Life Sci. Educ. **12**, 30 (2013).
[17] A. Miyake, L. E. Kost-Smith, N. D. Finkelstein, S. J. Pollock, G. L. Cohen, and T. A. Ito, Reducing the gender achievement gap in college science: A classroom study of values affirmation, Science **330**, 1234 (2010).
[18] R. H. Tai and P. M. Sadler, Gender differences in introductory undergraduate physics performance: University physics versus college physics in the USA, Int. J. Sci. Educ. **23**, 1017 (2001).
[19] 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).
[20] J. Docktor and K. Heller, Gender differences in both Force Concept Inventory and introductory physics performance, AIP Conf. Proc. **1064**, 15 (2008).
[21] L. McCullough, Gender, context, and physics assessment, J. Int. Women's Stud. **5**, 20 (2004), https://vc.bridgew.edu/jiws/.
[22] A. Traxler, R. Henderson, J. Stewart, G. Stewart, A. Papak, and R. Lindell, Gender fairness within the Force Concept Inventory, Phys. Rev. Phys. Educ. Res. **14**, 010103 (2018).
[23] M. J. Cahill, K. M. Hynes, R. Trousil, L. A. Brooks, M. A. McDaniel, M. Repice, J. Zhao, and R. F. Frey, Multiyear, multi-instructor evaluation of a large-class interactive-engagement curriculum, Phys. Rev. ST Phys. Educ. Res. **10**, 020101 (2014).
[24] 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).
[25] S. M. Aguillon, G.-F. Siegmund, R. H. Petipas, A. G. Drake, S. Cotner, and C. J. Ballen, Gender differences in

student participation in an active-learning classroom, CBE Life Sci. Educ. **19**, ar12 (2020).
[26] C. M. Steele, A threat in the air: How stereotypes shape intellectual identity and performance, Am. Psychol. **52**, 613 (1997).
[27] E. G. Bailey, R. F. Greenall, D. M. Baek, C. Morris, N. Nelson, T. M. Quirante, N. S. Rice, S. Rose, and K. R. Williams, Female in-class participation and performance increase with more female peers and/or a female instructor in life sciences courses, CBE Life Sci. Educ. **19**, ar30 (2020).
[28] S. L. Eddy, S. E. Brownell, P. Thummaphan, M.-C. Lan, and M. P. Wenderoth, Caution, student experience may vary: Social identities impact a student's experience in peer discussions, CBE Life Sci. Educ. **14**, 1 (2015).
[29] E. M. Nadile *et al.*, Call on me! Undergraduates' perceptions of voluntarily asking and answering questions in front of large-enrollment science classes, PLoS One **16**, e0243731 (2021).
[30] G. M. Walton and G. L. Cohen, A brief social-belonging intervention improves academic and health outcomes of minority students, Science **331**, 1447 (2011).
[31] G. M. Walton and G. L. Cohen, A question of belonging: race, social fit, and achievement, J. Pers. Soc. Psychol. **92**, 82 (2007).
[32] J. D. Edwards, R. S. Barthelemy, and R. F. Frey, Relationship between course-level social belonging (sense of belonging and belonging uncertainty) and academic performance in General Chemistry 1, J. Chem. Educ. (2021), https://doi.org/10.1021/acs.jchemed.1c00405.
[33] A. Fink, R. F. Frey, and E. D. Solomon, Belonging in general chemistry predicts first-year undergraduates' performance and attrition, Chem. Educ. Res. Pract. **21**, 1042 (2020).
[34] 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**, 1 (2018).
[35] S. Zumbrunn, C. McKim, E. Buhs, and L. R. Hawley, Support, belonging, motivation, and engagement in the college classroom: A mixed method study, Instr. Sci. **42**, 661 (2014).
[36] 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).
[37] C. J. Ballen *et al.*, Smaller classes promote equitable student participation in STEM, BioScience **69**, 669 (2019).
[38] J. R. Howard and A. L. Henney, Student participation and instructor gender in the mixed-age college classroom, J. Higher Educ. **69**, 384 (1998).
[39] L. M. Fritschner, Inside the undergraduate college classroom, J. Higher Educ. **71**, 342 (2000).
[40] J. Howard, A. Zoeller, and Y. Pratt, Students' race and participation in sociology classroom discussion: A preliminary investigation, J. Scholar. Teach. Learn. **6**, 14 (2006).
[41] G. Crombie, S. W. Pyke, N. Silverthorn, A. Jones, and S. Piccinin, Students' perceptions of their classroom participation and instructor as a function of gender and context, J. Higher Educ. **74**, 51 (2003).
[42] H. E. Tatum, B. M. Schwartz, P. A. Schimmoeller, and N. Perry, Classroom participation and student-faculty interactions: Does gender matter?, J. Higher Educ. **84**, 745 (2013).
[43] R. Cornelius, J. Gray, and A. Constantinople, Student-faculty interaction in the college classroom, J. Res. Dev. **23**, 189 (1990).
[44] J. C. Pearson and R. West, An initial investigation of the effects of gender on student questions in the classroom: Developing a descriptive base, Commun. Educ. **40**, 22 (1991).
[45] K. L. Brady and R. M. Eisler, Sex and gender in the college classroom: A quantitative analysis of faculty-student interactions and perceptions, J. Educ. Psychol. **91**, 127 (1999).
[46] M. Crawford and M. MacLeod, Gender in the college classroom: An assessment of the "chilly climate" for women, Sex Roles **23**, 101 (1990).
[47] S. L. Eddy and K. A. Hogan, Getting under the hood: How and for whom does increasing course structure work?, CBE Life Sci. Educ. **13**, 453 (2014).
[48] R. R. Weaver and J. Qi, Classroom organization and participation: College students' perceptions, J. Higher Educ. **76**, 570 (2005).
[49] D. Karp and W. Yoels, The college classroom: Some observations on the meanings of student participation, Sociol. Soc. Res. **60**, 421 (1976).
[50] E. Junn, "Pearls of Wisdom": Enhancing Student Class Participation with an Innovative Exercise, J. Instr. Psychol. **21**, 385 (1994).
[51] L. M. Daggett, Quantifying class participation, Nurse Educator **22**, 13 (1997).
[52] D. L. Garard, L. Lippert, S. K. Hunt, and S. T. Paynton, Alternatives to traditional instruction: Using games and simulations to increase student learning and motivation, Commun. Res. Reports. **15**, 36 (1998).
[53] J. A. Crone, Using panel debates to increase student involvement in the introductory sociology class, Teaching Sociology **25**, 214 (1997).
[54] C. Garside, Look who's talking: A comparison of lecture and group discussion teaching strategies in developing critical thinking skills, Commun. Educ. **45**, 212 (1996).
[55] D. G. Smith, College classroom interactions and critical thinking, J. Educ. Psychol. **69**, 180 (1977).
[56] R. Berdine, Why some students fail to participate in class, Marketing News **20**, 23 (1986).
[57] D. Dancer and P. Kamvounias, Student involvement in assessment: A project designed to assess class participation fairly and reliably, Assessment and Evaluation in Higher Education **30**, 445 (2005).
[58] M. Armstrong and D. Boud, Assessing Participation in Discussion: An exploration of the issues, Stud. Higher Educ. **8**, 33 (1983).
[59] K. Z. Girgin and D. D. Stevens, Bridging in-class participation with innovative instruction: use and implications in a Turkish university classroom, Innov. Educ. Teach. Int. **42**, 93 (2005).
[60] C. Gopinath, Alternatives to instructor assessment of class participation, J. Educ. Bus. **75**, 10 (1999).
[61] J. R. Howard, L. B. Short, and S. M. Clark, Students' participation in the mixed-age college classroom,

Teach. Sociol. **24**, 8 (1996). (accessed January 14, 2020).
[62] C. E. Nunn, Discussion in the college classroom: Triangulating observational and survey results, J. Higher Educ. **67**, 243 (1996).
[63] R. Frey, B. Fisher, E. Solomon, D. Leonard, J. Mutambuki, J. Luo, S. Pondugula, and C. Cohen, A Visual approach to helping instructors integrate, document, and refine active learning, J. Coll. Sci. Teach. **045**, 20 (2016).
[64] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.17.020140 for the detailed sample OPAL timelines, category definitions and additional segmenting guidelines, detailed statistical information, detailed study 2 findings, student surveys and gender participation map.
[65] J. Gess-Newsome, Pedagogical content knowledge: An introduction and orientation, in *Examining Pedagogical Content Knowledge* (Kluwer Academic Publishers, Dordrecht, 2006), pp. 3–17.
[66] L. S. Shulman, Those who understand: Knowledge growth in teaching, Educ. Res. **15**, 4 (1986).
[67] M. K. Smith, F. H. M. Jones, S. L. Gilbert, and C. E. Wieman, The classroom observation protocol for undergraduate stem (COPUS): A new instrument to characterize university STEM classroom practices, CBE Life Sci. Educ. **12**, 618 (2013).
[68] M. T. Hora, Exploring the Use of the Teaching Dimensions Observation Protocol to Develop Fine-grained Measures of Interactive Teaching in Undergraduate Science Classrooms (WCER Working Paper 2013-6), Madison (2013). https://wcer.wisc.edu/docs/working-papers/Working_Paper_No_2013_06.pdf (accessed October 19, 2021).
[69] E. D. Solomon, M. D. Repice, J. M. Mutambuki, D. A. Leonard, C. A. Cohen, J. Luo, and R. F. Frey, A mixed-methods investigation of clicker implementation styles in STEM, CBE Life Sci. Educ. **17**, ar30 (2018).
[70] S. B. Merriam, *Qualitative Research: A Guide to Design and Implementation* (Jossey-Bass, San Francisco, 2009).
[71] N. McDonald, S. Schoenebeck, and A. Forte, Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice, Proc. ACM Human-Computer Interact. **3**, 1 (2019).
[72] D. S. Kerby, The simple difference formula: An approach to teaching nonparametric correlation, Compr. Psychol. **3**, 11.IT.3.1 (2014), https://journals.sagepub.com/doi/full/10.2466/11.IT.3.1.
[73] C. O. Fritz, P. E. Morris, and J. J. Richler, Effect size estimates: current use, calculations, and interpretation, J. Exp. Psychol. Gen. **141**, 2 (2012).
[74] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. (L. Erlbaum Associates, Hillsdale, NJ, 1988).
[75] D. Z. Grunspan, S. L. Eddy, S. E. Brownell, B. L. Wiggins, A. J. Crowe, and S. M. Goodreau, Males under-estimate academic performance of their female peers in undergraduate biology classrooms, PLoS One **11**, e0148405 (2016).
[76] M. T. H. Chi and R. Wylie, The ICAP framework: Linking cognitive engagement to active learning outcomes, Educ. Psychol. **49**, 219 (2014).
[77] J. R. Howard and R. Baird, The consolidation of responsibility and students' definitions of situation in the mixed-age college classroom, J. Higher Educ. **71**, 700 (2000).
[78] J. R. Howard, G. H. James, and D. R. Taylor, The consolidation of responsibility in the mixed-age college classroom, Teaching Sociology **30**, 214 (2002).
[79] K. N. White, K. Vincent-Layton, and B. Villarreal, Equitable and inclusive practices designed to reduce equity gaps in undergraduate chemistry courses, J. Chem. Educ. **98**, 330 (2021).
[80] K. D. Tanner, Structure matters: Twenty-one teaching strategies to promote student engagement and cultivate classroom equity, CBE Life Sci. Educ. **12**, 322 (2013).
[81] J. E. Caldwell, Clickers in the large classroom: Current research and best-practice tips, CBE Life Sci. Educ. **6**, 9 (2007).